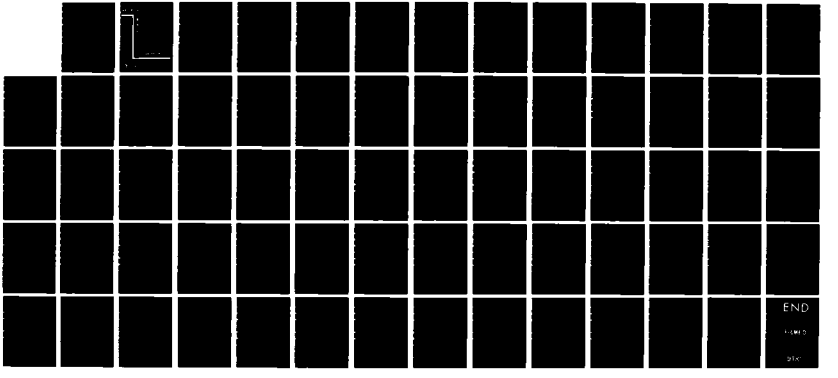
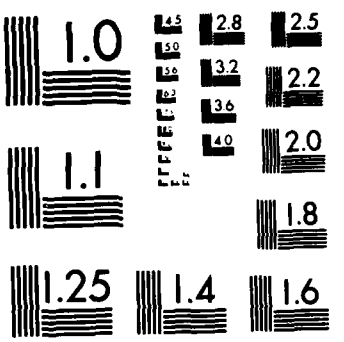
AD-A154 071 ESTIMATING ABILITY WITH THE WRONG MODEL(U) EDUCATIONAL 1/1
TESTING SERVICE PRINCETON NJ H WAINER ET AL. APR 85
AFHRL-TR-84-45 F41689-82-C-0020

UNCLASSIFIED F/G 12/1 NL





MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

AFHRL-TR-84-45

(2)

AIR FORCE HUMAN RESOURCES LABORATORY

ESTIMATING ABILITY WITH THE WRONG MODEL

By

Howard Wainer

Educational Testing Service
Princeton, New Jersey 08541

David Thissen

Department of Psychology
University of Kansas
426 Fraser Hall
Lawrence, Kansas 66045

MANPOWER AND PERSONNEL DIVISION
Brooks Air Force Base, Texas 78235-5601

April 1985
Final Report for Period March 1982 - September 1984

Approved for public release; distribution unlimited.

AD-A154 071

DTIC FILE COPY

DTIC ELECTE MAY 23 1985 S

B

AIR FORCE SYSTEMS COMMAND
BROOKS AIR FORCE BASE, TEXAS 78235-5000

NOTICE

When Government drawings, specifications, or other data are used for any purpose other than in connection with a definitely Government-related procurement, the United States Government incurs no responsibility or any obligation whatsoever. The fact that the Government may have formulated or in any way supplied the said drawings, specifications, or other data, is not to be regarded by implication, or otherwise in any manner construed, as licensing the holder, or any other person or corporation; or as conveying any rights or permission to manufacture, use, or sell any patented invention that may in any way be related thereto.

The Public Affairs Office has reviewed this report, and it is releasable to the National Technical Information Service, where it will be available to the general public, including foreign nationals.

This report has been reviewed and is approved for publication.

JAMES A. EARLES
Contract Monitor

NANCY GUINN, Technical Director
Manpower and Personnel Division

ANTHONY F. BRONZO, JR., Colonel, USAF
Commander

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE

REPORT DOCUMENTATION PAGE

1a. REPORT SECURITY CLASSIFICATION Unclassified	1b. RESTRICTIVE MARKINGS
2a. SECURITY CLASSIFICATION AUTHORITY	3. DISTRIBUTION/AVAILABILITY OF REPORT Approved for public release; distribution unlimited.
2b. DECLASSIFICATION/DOWNGRADING SCHEDULE	
4. PERFORMING ORGANIZATION REPORT NUMBER(S)	5. MONITORING ORGANIZATION REPORT NUMBER(S) AFHRL-TR-84-45
6a. NAME OF PERFORMING ORGANIZATION Educational Testing Service 6b. OFFICE SYMBOL (If applicable) T-21	7a. NAME OF MONITORING ORGANIZATION Manpower and Personnel Division
6c. ADDRESS (City, State and ZIP Code) Princeton, New Jersey 08541	7b. ADDRESS (City, State and ZIP Code) Air Force Human Resources Laboratory Brooks Air Force Base, Texas 78235-5601
8a. NAME OF FUNDING/SPONSORING ORGANIZATION Air Force Human Resources Laboratory 8b. OFFICE SYMBOL (If applicable) HQ AFHRL	9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER F41689-82-C-0020
8c. ADDRESS (City, State and ZIP Code) Brooks Air Force Base, Texas 78235-5601	10. SOURCE OF FUNDING NOS. PROGRAM ELEMENT NO. 62703F; PROJECT NO. 7719; TASK NO. 18; WORK UNIT NO. 27
11. TITLE (Include Security Classification) Estimating Ability with the Wrong Model	
12. PERSONAL AUTHOR(S) Wainer, Howard; Thissen, David	
13a. TYPE OF REPORT Final 13b. TIME COVERED FROM Mar 82 TO Sep 84	14. DATE OF REPORT (Yr., Mo., Day) April 1985 15. PAGE COUNT 65
16. SUPPLEMENTARY NOTATION	

17. COSATI CODES		
FIELD	GROUP	SUB. GR.
05	09	00
05	10	00

18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)
AMJACK estimator, Armed Services Vocational Aptitude Battery, Biweight estimator, H-estimates, and ~~item response theory~~ M-estimates.

19. ABSTRACT (Continue on reverse if necessary and identify by block number)

Using simulated item response data, the performance of several "robust" and conventional schemes for ability estimation was evaluated in conjunction with logistic item response theory models (one, two, and three parameter models). The simulated item response data were generated using a model that is more complex than are the usual logistic models; therefore, all three models were fundamentally (and realistically) "wrong." Consideration was given to estimation with a few responses (four) and with large numbers (20 and 40). With few item responses, the relative "wrongness" of the model had little effect, whereas the choice of estimator had more serious consequences. With many items, the choice of item response model made more difference than did the choice of estimator. Implications of these findings for computerized adaptive testing are discussed.

20. DISTRIBUTION/AVAILABILITY OF ABSTRACT UNCLASSIFIED/UNLIMITED ☒ SAME AS RPT. ☐ DTIC USERS ☐	21. ABSTRACT SECURITY CLASSIFICATION
22a. NAME OF RESPONSIBLE INDIVIDUAL Nancy A. Perrigo Chief, STINFO Office	22b. TELEPHONE NUMBER (Include Area Code) (512) 536-3877 22c. OFFICE SYMBOL AFHRL/TSR

DD FORM 1473, 83 APR EDITION OF 1 JAN 73 IS OBSOLETE.

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE

Item 18 (Continued)

robust estimation
simulation

TABLE OF CONTENTS

		Page
I	INTRODUCTION	3
II	THE LOGIC AND STRUCTURE OF THE SIMULATION	5
III	METHODS OF ABILITY ESTIMATION	9
IV	SCALING CONVENTIONS AND PERFORMANCE CRITERIA	16
V	THE TESTS	20
VI	RESULTS, NATURAL SCALING	32
VII	RESULTS, WITH RESCALING	44
VIII	CONCLUSIONS	49

LIST OF TABLES

TABLE		Page
1	Fitting Four ASVAB Items With Three Models	7
2	Multiple Category and *n*-PL Parameters, WK 4-Item Set	23
3	Multiple Category n-PL Parameters, GS 4-Item Set	26
4	Multiple Category and n-PL Parameters for 20-Item Set	28

LIST OF ILLUSTRATIONS

DISPLAY		Page
1	An Ideal Version of Item Characteristic Curve With Item Parameters Illustrated	3
2a	Item Characteristic Curves for Four ASVAB Items Fitted With the Multiple Category Model	8
2b	The Fitted ICCs From the 1-PL Shown Against the True ICCs	10
2c	The Fitted ICCs From the 2-PL Shown Against the True ICCs	11
2d	The Fitted ICCs From the 3-PL Shown Against the True ICCs	12
3	"True Trace Lines" and 3-PL Estimates of the Four-Item WK--Inspired Example	22
4	"True Trace Lines" and 3-PL Estimates of the Four-Item GS--Inspired Example	25

DISPLAY Page

5 XTREE Plot Showing Squared Bias and Variance for the Simulated 4-Item Word Knowledge Test With RMSE Rescaling 34

6 XTREE Plot Showing Squared Bias and Variance for the Simulated 4-Item General Science Test With RMSE Rescaling 35

7 Frequency Distributions of Bias and Standard Error for the AMJK Estimator From the Simulated 4-Item Word Knowledge Test 36

8 Frequency Distributions of Bias and Standard Error for the 3-PL Mode Estimator From the Simulated 4-Item Word Knowledge Test 37

9 XTREE Plot Showing Squared Bias and Variance for the Simulated 20-Item General Science Test With RMSE Rescaling 39

10 XTREE Plot Showing Squared Bias and Variance for the Simulated 40-Item General Science Test With RMSE Rescaling 40

11 Frequency Distributions of Bias and Standard Error for the 3-PL Mean Estimator From the Simulated 20-Item General Science Test 41

12 Frequency Distributions of Bias and Standard Error for the 3-PL Mean Estimator From the Simulated 40-Item General Science Test 42

13 Frequency Distributions of Bias and Standard Error for the 3-PL Mode Estimator From the Simulated 20-Item General Science Test 43

14 XTREE Plot Showing Squared Bias and Variance for the Simulated 4-Item Word Knowledge Test With REVM Rescaling 45

15 XTREE Plot Showing Squared Bias and Variance for the Simulated 4-Item Word Knowledge Test With REMI Rescaling 46

16 XTREE Plot Showing Squared Bias and Variance for the Simulated 4-Item General Science Test With REVM Rescaling 47

17 XTREE Plot Showing Squared Bias and Variance for the Simulated 4-Item General Science Test With REMI Rescaling 48

DISPLAY		Page
18	XTREE Plot Showing Squared Bias and Variance for the Simulated 20-Item General Science Test With REVM Rescaling	50
19	XTREE Plot Showing Squared Bias and Variance for the Simulated 20-Item General Science Test With REMI Rescaling	51
20	XTREE Plot Showing Squared Bias and Variance for the Simulated 40-Item General Science Test With REVM Rescaling	52
21	XTREE Plot Showing Squared Bias and Variance for the Simulated 40-Item General Science Test With REMI Rescaling	53

Accession For
NTIS GRA&I ☑
DTIC TAB ☐
Unannounced ☐
Justification
Distribution/
Availability Codes
Dist | Avail and/or Special
A-1

SUMMARY

No model is ever a perfect reflection of the data it is built to summarize. There are always errors of fit. This is as true with modern item response theory (IRT) as with all other models. It is important to know to what extent the accuracy of measurement made with these models is perturbed by misfit and what can be done to minimize the inaccuracy.

First, a detailed general model was fit to ASVAB (Armed Services Vocational Aptitude Battery) data to provide the framework for a realistic simulation structure. Then three of the most commonly used IRT models were fit in this simulation. A variety of robust estimators of ability were used and the accuracy and efficiency of each estimator was determined.

With short tests, a simple model coupled with a robust estimator seemed to be the methodology of choice for describing the data. As test length increased, so too did the benefits of utilizing a more complex parameterization.

An unexpected finding was that coupling robust estimators with a Bayesian prior yielded substantial shrinkage. Future work on ability estimation, especially for practical applications of adaptive testing, is required to "unshrink" ability estimates.

PREFACE

This study was conducted under Project 7719, Force Acquisition and Distribution Systems. The effort represents the concern of this Laboratory for advancing the state-of-art in applications of item response theory to selection testing.

ACKNOWLEDGEMENT

The authors would like to thank James Earles of AFHRL, Neil Dorans and Marilyn Wingersky of ETS for helpful comments.

ESTIMATING ABILITY WITH THE WRONG MODEL

I. Introduction

"All theories are wrong. It's just that some are easier to disprove than others."

(John W. Tukey)

A model is never a perfect mirror of reality. It is a simplification that has useful properties. The mathematical models that constitute modern item response theory (IRT) vary in their complexity, but all are simplifications. This study examines the extent to which these simplifications disturb the accuracy of the estimation of ability and the extent to which accuracy can be affected by the choice of the estimation algorithm.

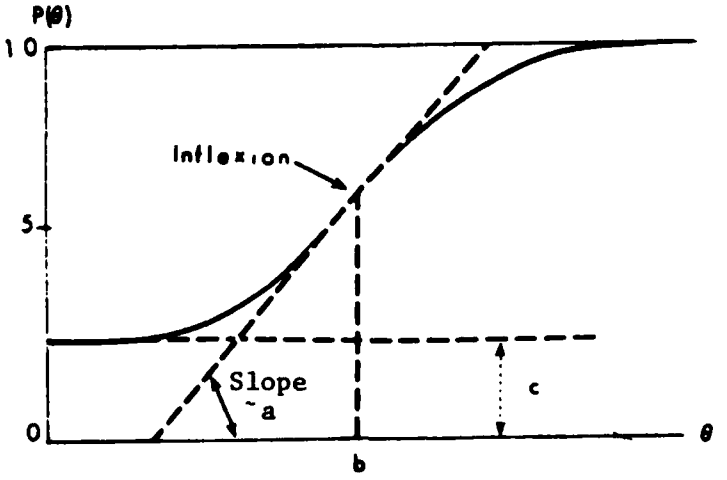
We shall confine our investigation to three of the most popular IRT models. These are the one-, two-, and three-parameter logistic models (denoted 1-PL, 2-PL and 3-PL). Their mathematical form is depicted in equations (1), (2) and (3), respectively. Details about their genesis and explanations and justifications of their use are found in the standard sources (e.g. Allen & Yen, 1979; Andersen, 1980; Lord, 1980; Lord & Novick, 1968; Rasch, 1960; Wright & Stone, 1979)

$$P(\theta) = 1/(1 + \exp(-1(\theta - b))) \quad (1)$$

$$P(\theta) = 1/(1 + \exp(-a(\theta - b))) \quad (2)$$

$$P(\theta) = c + (1-c)/(1 + \exp(-a(\theta - b))) \quad (3)$$

$P(\theta)$ is the probability of choosing the correct answer expressed as a function of the person's ability θ; b is the item difficulty expressed in the same metric as ability (logits); a is proportional to the slope of the item characteristic curve at its steepest point, and c is the lower asymptote of the item characteristic curve. This latter parameter is an attempt to model the effects of guessing, in that when someone's ability is well below the difficulty of the item, it is assumed that plausible behavior is random guessing among all the alternatives offered. A graphic interpretation of these parameters is shown in Display 1 (from Lord, 1980, p. 14).



Display 1. An ideal version of item characteristic curve (from Lord, 1980, p. 14, reprinted with permission) with item parameters illustrated.

Each model is obviously a generalization of the one before it; we can easily specialize the 3-PL to the 2-PL and the 2-PL to the 1-PL. As we shall describe later, there are costs to be borne through the use of an overly general model; so it is wise to use the most restricted model that represents the data at hand sufficiently well.

I.a Which ability?

In this study we are concerned with the estimation of ability — nothing more. When looking at equations (1), (2) and (3) one can be misled into thinking that all three of the θ's are the same. There is a sense in which this is not true. In models (1) and (2), the ability parameter is the same, but in (3) it is not. For example, let

$$P = .5 \quad a = 1 \quad b = 0 \text{ and } c = .2$$

We then find that

		1-PL	2-PL	3-PL
θ	equals	0	0	-.5

A glib explanation of this phenomenon is that, if the model allows guessing, less ability is required to get the item correct. In this example we see that about half a logit less ability is required. The question that naturally arises after noticing this is, "Which θ are we trying to estimate?" The answer is, "The right one." More about this later.

I.b Which model?

Previous studies of the usefulness of item response models have taken one of two approaches:

1. Fit various models to real data and see which one fits best. This approach has the enviable property that the investigator knows the model is being tested under realistic conditions. The principal drawback is that one does not know what is the right answer. Thus we may discover that a particular model provides the best answer, but we do not know if the best we can do is good enough. Or, under a more optimistic viewpoint, if a less general model is still good enough for the purposes at hand.

2. Fit various models to simulated data. This approach has the advantage of allowing the investigator to know the correct answer, but has the drawback of having an uncertain relationship to reality. It also tends to trivialize the study, for almost surely the model that is used to generate the data will be the winner in any competition.

In this study we have tried to solve these two problems simultaneously. We do this by using the most realistic model we could find to generate the data for the simulation. Details about this are in section II.

I.c Which estimator?

The structure of our study is starting to emerge. We will fit real data with a model too complex and cumbersome for common use, but which is glovelike in its matching of the observed data. We then use the item parameters of this model to provide a parameterized version of the trace lines. This model also provides the right value of ability which we shall try to recover. We then fit the three logistic models to these data. Each model fits imperfectly. Last, we use these imperfect models to estimate ability. Naturally we expect them all to fall short. The interesting questions are "how badly do they do?" and "can we 'fiddle' with the estimator to improve its performance?" In this study we 'fiddle' with estimators in a variety of ways, leaning heavily on theoretical and practical developments in the robust/resistant estimation area. We shall test many kinds of estimators. These are described in section III.

I.d Who won?

As is often the case in contests like this, there is no clear-cut answer to what is best; although some clear favorites do emerge. There are some obvious losers. Our findings are quite optimistic. Earlier work (Thissen & Wainer, 1982) indicated that the 3-PL required enormous norming samples for accurate estimates of item parameters for items of only modest difficulty. Later work (Jones, Wainer, & Kaplan, 1984) demonstrates how inaccuracies in the estimation of the item parameters become error in the ability estimates. These results, coupled with the empirical knowledge that guessing does take place, had led us toward doubts about the ease with which item response models could be used accurately in many practical situations. These doubts have been assuaged, for we have discovered that the bias introduced through the use of the wrong response model is smallish and can be corrected somewhat with robust estimation wizardry.

The themes and broad outcomes thus sketched, let us now go into the myriad of detail required for a deep understanding of our study and its implications.

II. The Logic and Structure of the Simulation

We wished to conduct a simulation whose structure, although known to us, would be inextricably tied to the empirical world. To accomplish this we fit data from the Armed Services Vocational Aptitude Battery (ASVAB) with an item response model that is so general that even non-monotonicities in the item response curve could be fit easily. The IRT model used was the "multiple category model" described by Thissen and Steinberg (1984). As the name implies, it is a model which has a distinct trace line for each of the

So for estimators with "natural" variance near two, β is near unity; if such an estimator is linearly related to u_j, UBAR(j,k) is near u_j and $\sum_j u_j$UBAR(j,k) is near 10 and β' is also near unity. For a "perfect" estimator, all three scalings are identical.

For less accurate estimators, the rescalings may or may not diverge. Variance matching changes the scale of the estimates (usually, but not always, it expands them) to match the variance of the original; MIN(RMSE) rescaling may not match the variance. The most extreme divergence comes with a hypothetical estimator which is uncorrelated with the generating u_j's. Variance matching rescaling would produce a set of estimates with variance 2, all of which would be error. MIN(RMSE) rescaling, on the other hand, would rescale such a set of estimates to have variance zero (β' would be zero).

IV.d. What is Which Rescaling For?

All of the model/estimator combinations are supposed to be estimating the same thing. Therefore, one might imagine that a testing program using one estimator could "switch" to another and maintain comparability between u-estimates through the "inherent equating" of the IRT model. The "naturally" scaled versions of the 28 estimators show what would happen if such a switch were made. In general, the results would be disastrous. The different scales, resulting mostly from different degrees of shrinkage, make the results using most of the pairs of estimators hopelessly incomparable.

Shrinkage seems to be fairly uniform within an estimator, however. And the scale of IRT estimators is usually arbitrarily set to have a predetermined variance in some standardizing sample. If the variance of the population is known (as it is in our simulation) or if there exists some standardizing sample in which the variance is assumed known, the output of an estimator, however much it "shrinks," may be rescaled to have that variance. We have done this linearly and call this "variance matching" (VM) rescaling. Lord (1984) has pointed out that the shrinkage to be removed by rescaling is not linear, and so linear rescaling is not a perfect solution. We expect future developments in IRT to include some more elegant techniques for "unshrinking." No others are readily available at this time.

In any event, VM gives the results that would be obtained if the estimator was used in a form rescaled to have the variance of some standardizing sample. It "removes" the artifactual differences due to differential shrinkage. With "variance matching," we compare 28 model/estimator combinations, all of which give the same variance for our sample of simulees, on bias and random variance at each of the five levels of u. Given that the scale is arbitrary, this rescaling gives the performance for each estimator in contexts in which absolute cut-point values are used for admission, etc. relative to known standardizing distribution.

IV.b. Rescaling to Minimize MSE

One might also be concerned with the linear transformation of the estimated UHAT' = $\alpha' + \beta' \cdot$UHAT which gives the "best" (MSE) linear relationship with the five values of u_j (-2, -1, 0, 1, 2). [Actually, there are an infinite number of weighted versions, but we will restrict ourselves to the equally-weighted one.] For this, we want β' and α' to minimize

$$Q = \sum_j \{[(\beta' \text{UBAR}(j,k)+\alpha') - u_j]^2 + \beta'^2 \text{S.D.}(j,k)^2\} .$$

To minimize Q,

$$\frac{\partial Q}{\partial \alpha} = 2\sum_j \{[\beta' \text{UBAR}(j,k)+\alpha'] - u_j\} = 0$$

at the solution. Therefore

$$\alpha' = -\sum_j \beta' \text{UBAR}(j,k)/5 \quad .$$

$$\frac{\partial Q}{\partial \beta'} = 2\sum_j \{[(\beta' \text{UBAR}(j,k)+\alpha') - u_j]\text{UBAR}(j,k) + \beta' \text{S.D.}(j,k)^2\} = 0$$

at the solution. Therefore,

$$\beta' = \sum_j u_j \text{UBAR}(j,k) / \{\sum_j [(\text{UBAR}(j,k)^2 - \text{UBAR}(.,k)^2) + \text{S.D.}(j,k)^2]\}$$

The rescaled UHAT' = $\alpha' + \beta'$UHAT, giving minimum (MIN) RMSE produces three more indices of the quality of the estimators:

$$\text{DBMI}(j,k) = u_j - \text{UHAT}'(j,k)$$

$$\text{SDMI}(j,k) = \text{Standard Deviation}[\text{UHAT}'(j,k)]$$

and

$$\text{REMI}(j,k) = \text{SQRT}[\text{DBMI}(j,k)^2 + \text{SDMI}(j,k)^2] \ .$$

IV.c. A Note On the Rescalings

For highly accurate estimators, the three scalings are approximately the same. If the mean of the estimates is zero, $\alpha = \alpha' = 0$. And

$$\beta' = \beta^2 \{\sum_j u_j \text{UBAR}(j,k)/10\} \ .$$

IV.a. Rescaling to Equal Variance

For ease in interpretation, we would like to keep the integral values of the generating u_j's (-2, -1, 0, 1, 2). So we have chosen to leave the generating distribution "as it is" and "pseudo-standardize" the distributions of the estimates for each estimator by "matching" the mean and variance to that of the generating u's. The mean of (-2, -1, 0, 1, 2) is zero and its variance is two. So we require a linear transformation $UHAT^* = \alpha + \beta \cdot UHAT$ with mean zero and variance two.

This requires

$$\alpha = -\sum_j \beta \cdot UBAR(j,k)/5$$

as well as

$$Variance[UHAT^*(k)] = 0.2 \sum_j [\beta^2 (UBAR(j,k)^2 - UBAR(.,k)^2) + \beta^2 S.D.(j,k)^2]$$

Additionally,

$$Variance[UHAT^*(k)] = 2.$$

Combining gives

$$\beta^2 \sum_j [(UBAR(j,k)^2 - UBAR(.,k)^2) + S.D.(j,k)^2] = 10$$

so

$$\beta = SQRT(10/\{\sum_j [(UBAR(j,k)^2 - UBAR(.,k)^2) + S.D.(j,k)^2]\})$$

The rescaled $UHAT^* = \alpha + \beta \cdot UHAT$, in which α and β are as given above, are called "variance matched" (VM) because a set of them has the same mean (zero) and variance (2) as the generating distribution of five points. The UHAT* give three more criteria for the quality of estimator k:

$$DBVM(j,k) = u_j - UHAT^*(j,k)$$

$$SDVM(j,k) = Standard\ Deviation[UHAT^*(j,k)]$$

and

$$REVM(j,k) = SQRT[DBVM(j,k)^2 + SDVM(j,k)^2] \quad .$$

We have accumulated a total of 28 ability estimators. Nine estimators (MEAN, M.25, M.50, M1.0, MODE, H0.5, H1.0, H2.0 and BIWT) are defined and easily computable for all three logistic models, giving a total of 27. And AMJK is used only with the 1-PL model, for 28. Thissen, Wainer and Rubin (1984) describe a computer program for the simulation described here, in which subroutines compute each of the 28 estimators defined; those subroutines constitute another, more concrete, if less readable, definition of the estimators. For quick reference verbal descriptions of the estimators with their identifying codes are given in the glossary of this report.

Since all estimators included a N(0,1) population density as a "prior," it is useful for some purposes to describe the performance of the estimators in standard deviation units of that population. We follow this practice throughout this report.

IV. Scaling Conventions and Performance Criteria

In this simulation, the distribution of the values of ability (hereafter called "u" instead of θ to match the computer output which makes up much of the rest of the report) used to create the item responses is known: the distribution is 20% (usually 100) at each of five points $\{u_j = -2, -1, 0, 1, 2$, for j=1 to 5$\}$. Each model/estimator combination (k=1,28; e.g. 1-PL/MEAN, 3-PL/MODE, etc.) produces a mean value UBAR(j,k) for data for each of the five values of u_j. There is also a corresponding standard deviation (S.D.(j,k)) for each distribution of estimates within the set of data for a constant value of u_j. These values give three indices of the quality of estimator k at u_j:

$$\text{DBAR}(j,k) = u_j - \text{UBAR}(j,k)$$

$$\text{S.D.}(j,k) = \text{Standard Deviation [UHAT}(j,k)]$$

and

$$\text{RMSE}(j,k) = \text{SQRT}[\text{DBAR}(j,k)^2 + \text{S.D.}(j,k)^2] \ .$$

These three indices, of bias, random error, and mean squared error (MSE) respectively, reflect the quality (smaller=better) of each estimator as an estimate of the generating u_j.

However, it is well-known that IRT ability estimates are determined only up to a linear transformation of scale. And the several estimators to be compared here use different scales. Some of the more "robust" estimators "reject" more of the "information" in the data, and "regress" more back toward the mean (zero) — leaving them on a scale with different units. A scale-free comparison of the estimators is available if all are transformed to have the same mean and variance in some sample distribution.

M-estimator. Specifically, if $\hat{\theta}_{all}$ is the MODE for a particular person's response vector, we define

$\hat{\theta}_{(j)}$ = the MODE for that person with the $j^{\underline{th}}$ item omitted.

$\hat{\theta}^{*}_{(j)} = n\,\hat{\theta}_{all} - (n-1)\,\hat{\theta}_{(j)}$ is the jackknifed pseudo-value for the $j^{\underline{th}}$ item.

Next we form two vectors

$\underline{\hat{\theta}}^{*}_{c}$ of all values of $\hat{\theta}^{*}_{(i)}$ for which $x_i = 1$

$\underline{\hat{\theta}}^{*}_{w}$ of all values of $\hat{\theta}_{(i)}$ for which $x_i = 0$.

Then $\underline{\hat{\theta}}^{*}_{c}$ is of length $r = \sum_{i=1}^{h} x_i$, and

$\underline{\hat{\theta}}^{*}_{w}$ is of length $n-r$.

Let

$$\hat{\theta}_{c} = \text{AMT}\,(\underline{\hat{\theta}}^{*}_{c})$$

$$\hat{\theta}_{w} = \text{AMT}\,(\underline{\hat{\theta}}^{*}_{w})$$

where AMT(•) is the Sine M-estimate of the location of vector (•). Then the final AMJACK estimate is

$$\text{AMJACK} = [\,r\,\hat{\theta}_{c} + (n-r)\,\hat{\theta}_{w}]/n \qquad (10)$$

The AMJACK estimator has been shown to provide a superefficient estimator for the 1-PL compared to the MODE when sample sizes are small, even when the data conform to the assumptions that would make the MODE asymptotically optimal. Further, we have found that when the data deviate from the model, the AMJACK still maintains reasonable efficiency. Also, despite its arcane derivation, it is easily calculated for the 1-PL. For more general models, calculations begin to get heavy, but for most purposes this is not a serious problem. The two parts of the AMJACK work in concert to improve estimation. The Jackknife portion provides a direct estimate of estimator variability as well as reducing estimator bias. The Sine M-estimation part of AMJACK reduces the effects of unusual observations and so corrects for unmodelled guessing or "sleeping."

Comparing the h-estimator with the MODE, consider the 2-PL. $P' = aPQ$; thus the MODE is the solution to $(x-P)P'/PQ = (x-P)a = 0$. If we let h=0, the h-estimate is the solution to $(x-p)a\ (PQ)^0 = (x-P)a = 0$. Thus, the h-estimate for the 2-PL is identical to the MODE when h=0 and gets steadily more robust as h increases. In this study we shall use a variety of values of h. The estimates are labeled H0.5, H1.0, and H2.0 for h=0.5, 1.0, and 2.0, respectively. For further details see Jones' (1982a) development.

c. Biweight (BIWT) — Mislevy and Bock (1982) suggested this estimator in which the weighting function is

$$w(T) = \begin{cases} (1-U^2)^2 & \text{for } |U| \leq 1 \\ 0 & \text{otherwise} \end{cases} \qquad (9)$$

where

$$U = \frac{a\,(b - T)}{k}$$

Thus, we see that the weight is largest when b=T, i.e., when the difficulty of the item answered is the same as the examinee's ability. As the difference between b and T increases (as the item becomes less and less appropriate), w(T) decreases until the item is more than k units away, when it has no effect on the score at all. Mislevy and Bock suggest that a value of k=4 be used. In all of our tests we conform to their advice.

In Jones (1982b) is a comparison of the influence functions for these three M-estimators. The general shape of them is the same in our application, although they are shrunken a bit through the use of the N(0,1) prior.

It must be remembered that for all three of the M-estimators we have included a prior distribution, which, in each case, is simply the addition of one more term (G'/G). In all comparisons, the same prior is used for all estimators.

2. EAP estimators (MEAN and Mnnn) — Expected a posteriori estimators are another class of estimators. They have basically the same structure as the M-estimators except they reflect the mean of the posterior density, rather than the mode. When the prior is correct, EAP estimators are optimal, if mean square error is used as the criterion. We tested EAP analogs to the three h-estimates (above), called M.25, M.50, and M1.0. Bock (1983) provides a description of EAP estimation in an IRT context.

3. AMJACK estimation (AMJK) — This estimator, developed by Wainer and Wright (1980) is a combination of M- and L-estimates. It is made up from order statistics derived from the jackknifed pseudovalues of the modal ability estimate, which are then subjected to another

The results we shall report are optimistic to the extent that one's knowledge of the prior is incorrect. The estimators we used were

1. M-estimators — these estimators (here denoted T) are a general class that satisfy the equation

$$\Sigma\ w(T)\ [x - P(T)] = 0 \qquad (5)$$

for some weighting function w. Three of the estimators we used fall into this class. They are

a. MODE — The maximum likelihood estimator. This is the traditional estimator used to estimate ability. In this situation the weighting function is

$$w(T) = P'/(PQ), \qquad (6)$$

where Q = 1-P.

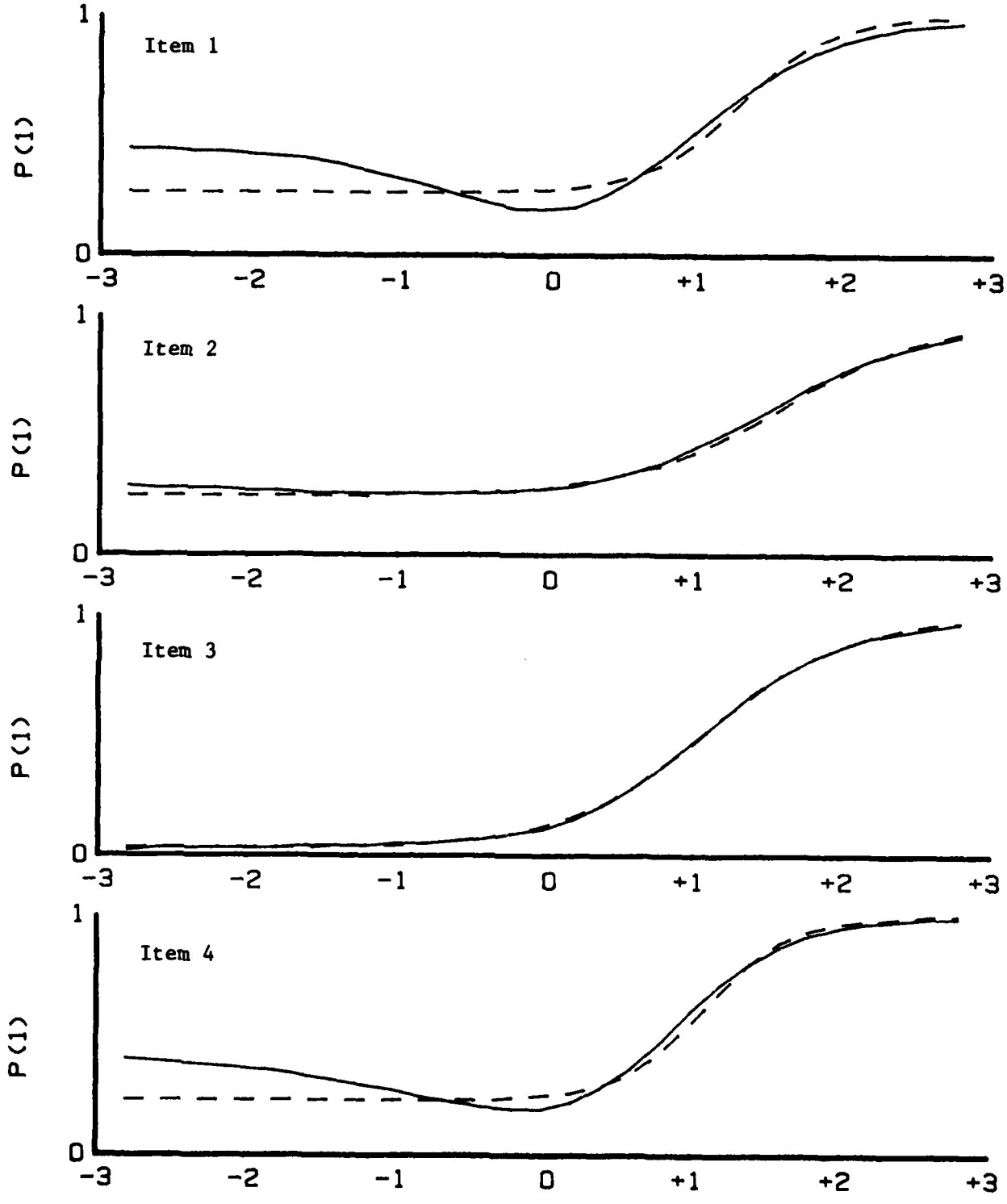
The solution to equation (5) with w as in equation (6) is the mode of the posterior density distribution. In our application we used a prior distribution as well. This corresponds to adding one more term to the sum in equation (5). If we denote the prior distribution G (Θ), this added term is G'/G. Thus, the MODE with prior distribution G is the solution T to the equation

$$\Sigma[\{P'/PQ\}\ (x-P)] + G'/G = 0 \qquad (7)$$

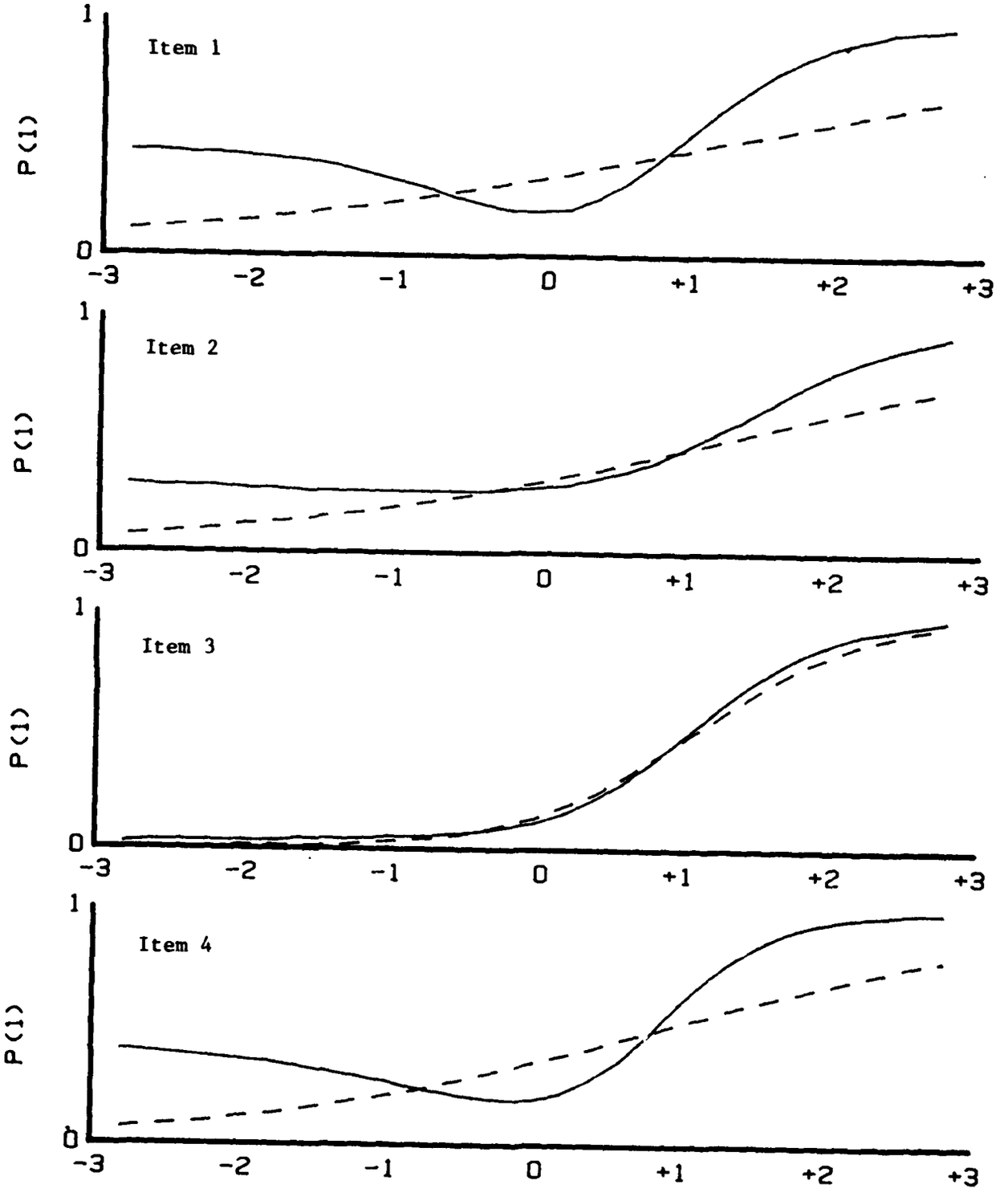
b) h-estimator (Hnnn) — These are a family of M-estimates developed by Jones (1982a), in which the weighting function is

$$w(T) = a\ (PQ)^{h} \qquad (8)$$

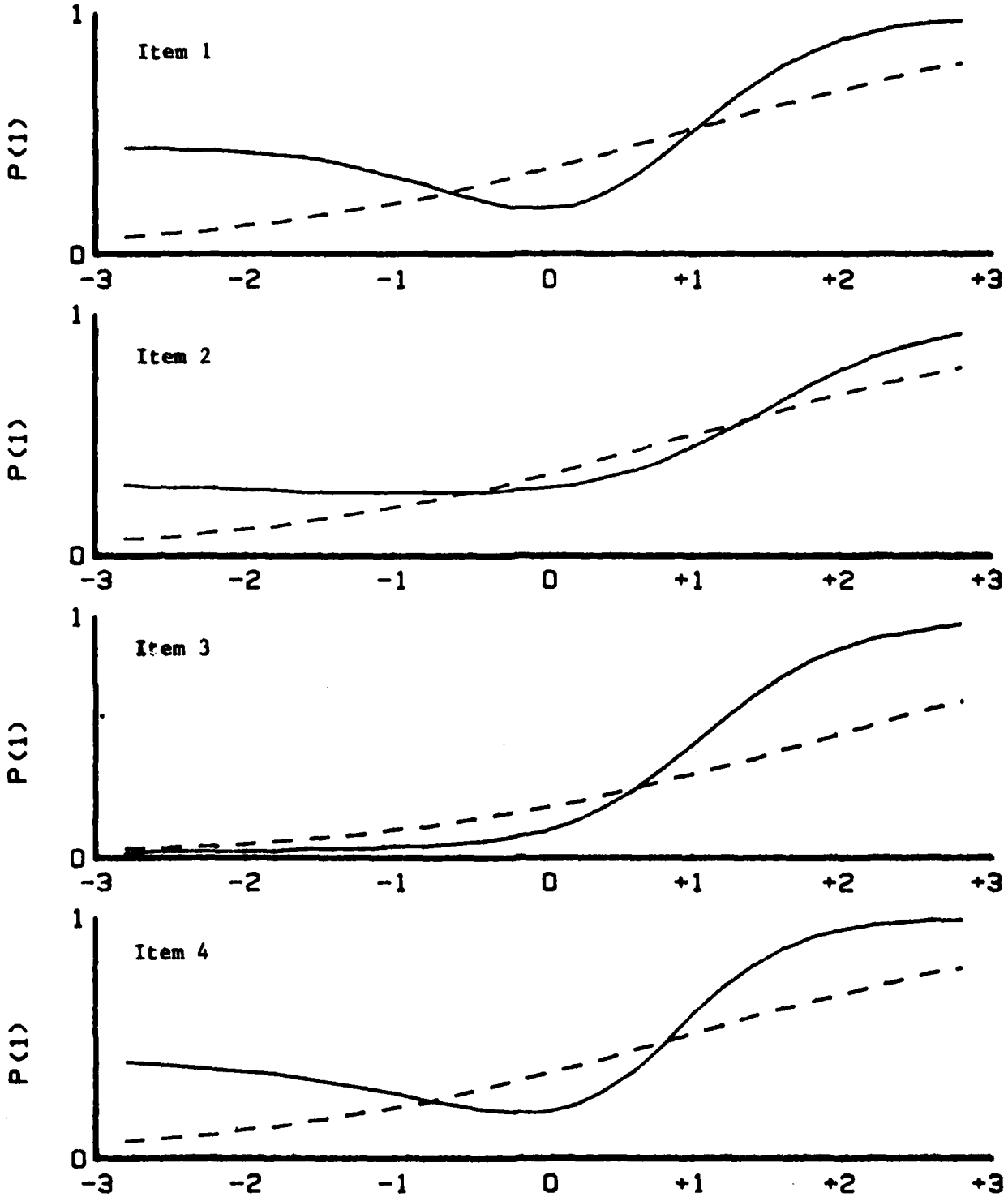
A few observations on the structure of this function, and its relationship to h can provide insights as to its robustifying character. First note that by weighting by the slope parameter, items that are more sharply discriminating count more toward the estimation of ability than do less discriminating items. The factor PQ is of greater importance if h is greater than one. This factor causes items near the examinee's ability (when P is most nearly equal to Q) to have the greatest weight, and items farther from that point to decrease in influence. Thus, items that are, say, too difficult for the examinee, and hence prime candidates for a guessing strategy, are downweighted. The extent of this downweighting depends on the value of h. The larger h, the more drastic the downweighting.



Display 2d. The fitted ICCs (dashed) from the 3-PL shown against the true (solid) ICCs.



Display 2c. The fitted ICCs (dashed from the 2-PL shown against the true (solid) ICCs.



Display 2b. The fitted ICCs (dashed) from the 1-PL shown against the true (solid) ICCs.

An examination of the different interpretations given to the results by the three models provides insights into these models. For example, item 3 is seen as the most difficult by the 1-PL (b=1.9), whereas it is the second easiest for the other two models (both agree on b=1.1). The explanation is clear, for the 1-PL has no way to deal with the substantial guessing seen on the other items except to call them easy (a lot of people got them correct); thus, the one item that has little guessing looks harder. The 2-PL deals with the guessing by making the slopes gradual. ("There is no guessing, it is just that it takes a long while for the curve to reach zero".) Thus, the 2-PL does better on the location parameter but seems to 'flub' on the slopes. The 3-PL does a better job all around but still has some misfit. The fitted curves for these four items for the three models are shown in Displays 2b, 2c, and 2d.

It is important to keep in mind that our method of fitting the ICCs by the three models essentially assumes that the norming sample is infinite. That is, the points used to estimate the parameters are assumed known without error. We will deal with the implications of this later; for now it tilts the contest in favor of the 3-PL which needs large samples to obtain stable parameter estimates.

II.b Generating Response Patterns

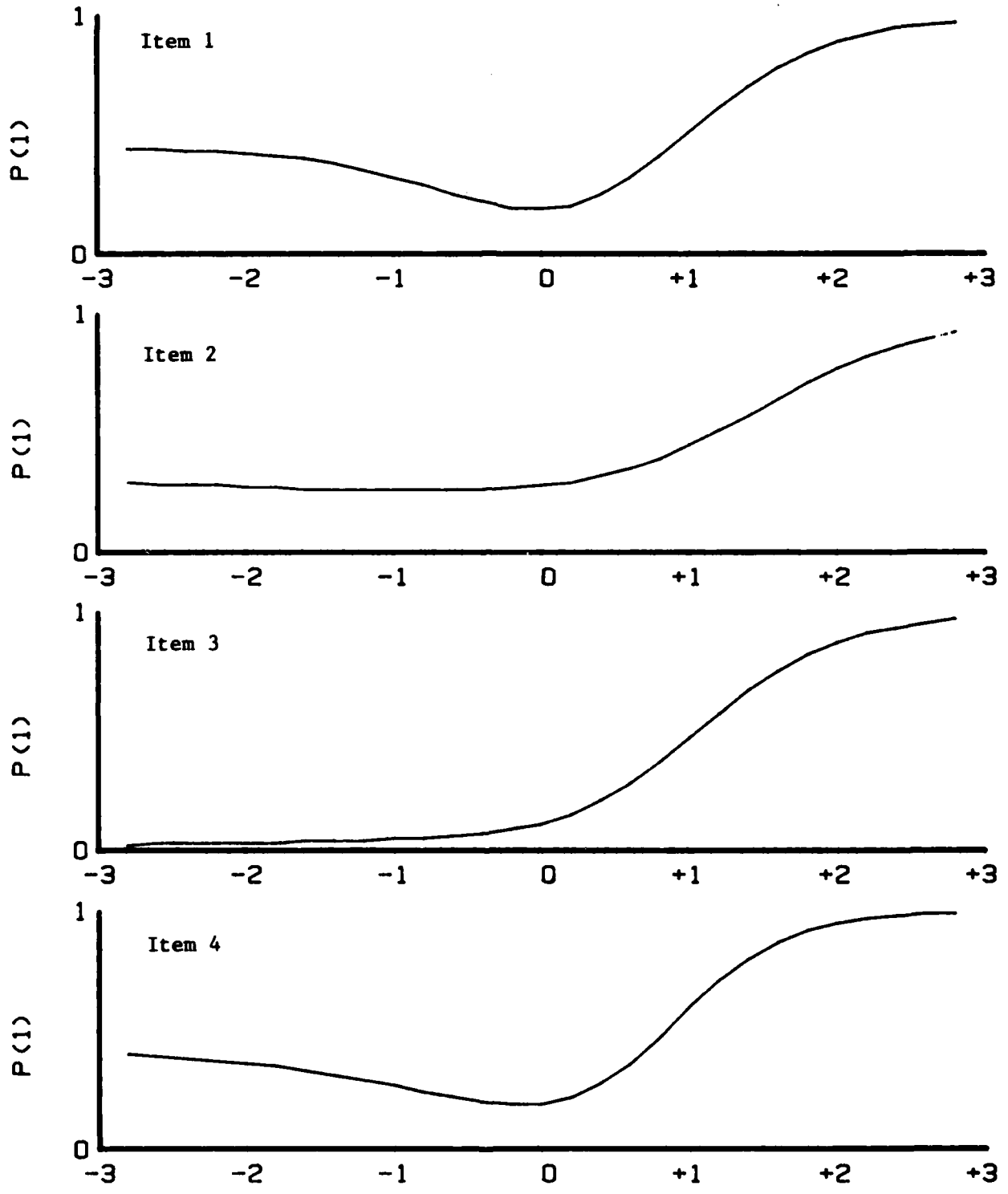
The next stage of the study involves generating 100 response patterns for the four items for each of five values of θ (-2, -1, 0, 1, and 2). These response patterns are generated from the "true model". When this is done, we then estimate θ under each of the assumed models, whose parameters are shown in Table 1. We then compare the accuracy of the estimates of θ for each model with the true values. The schemes used to estimate ability are described in detail in the next section.

III. Methods of Ability Estimation

There are no "straw men" in this study. We tried to use only those methods of estimation that are generally believed to be sensible. All estimators are Bayesian in the sense that we use a N(0,1) population distribution, included as a "prior" in estimation of θ. This was done for two reasons:

1. With a four-item "test" there would be many score vectors of (0,0,0,0) or (1,1,1,1). Without consideration of the population distribution, these would yield infinite ability estimates.

2. The best of current practice seems to be leaning heavily toward this approach.



Display 2a. Item characteristic curves for four ASVAB items fitted with the multiple category model (Thissen & Steinberg, 1984)

about the structure of the estimators. Second, we could ennumerate the entire space of 2^4 possible reponse patterns and compare the model generated posterior distributions with those actually observed. Third, exposition is eased considerably with the possibility of guiding the reader through an example of modest proportions. Fourth, but by no means least, a four item test is a realistic length for some purposes. Specifically, in adaptive tests an ability estimate is required after each response; thus, it is wise for us to know how accurate the estimate is after the first four items. Also, in many multifaceted tests, specific traits are tested with only a few items.

Of course our four-item test is but the beginning of this study. After we have gone carefully through our results for this situation, we shall describe the results for tests of various lengths, pointing out the differences in conclusions, such as they are, which depend on test length.

Thissen and Steinberg (1984) fitted the multiple category model to data on a subtest of the ASVAB. Display 2a includes the item characteristic curves for the four most difficult items. For several reasons these four were chosen to start our investigation. Principal among them is the fact that the ASVAB is not a very difficult test, and so many of the ICCs were not interesting (being just horizontal lines at the upper asymptote). Three of these ICC's show an interesting non-monotonicity at the very low end. The next step in our study was to fit the three logistic test models (equations (1), (2) and (3)) to these curves through maximum likelihood. We did this by choosing a large number of points (31) on these curves and fitting with the assumption that the ability distribution is Gaussian with mean zero and variance one ($\theta \sim N(0,1,)$). The results of this fitting are shown in Table 1.

Table 1. Fitting Four ASVAB Items with Three Models

Item	1-PL	2-PL	3-PL	
1	.93	1.36	1.30	
2	1.00	1.37	1.56	Locations
3	1.90	1.10	1.10	
4	.87	.88	1.08	
1	.73	.47	3.25	
2	.73	.51	1.94	Slopes
3	.73	1.68	2.14	
4	.73	.72	3.50	
1	0	0	.26	
2	0	0	.25	Lower Asymptotes
3	0	0	.03	
4	0	0	.23	

response alternatives for a multiple choice item. We are concerned here only with the trace line fitted to the correct response; but by using information in all of the response alternative data, the multiple category model may yield a wide variety of shapes for the probability of choosing the correct alternative as a function of ability. Specifically, the trace line for the correct response may be non-monotonic and does not necessarily have the extensive symmetry that is present in the conventional logistic models.

In the notation used by Thissen and Steinberg, the model for the correct alternative is

$$P(\text{correct}|\theta) = \frac{\exp(z_{\text{correct}}) + d_{\text{correct}}[\exp(z_0)]}{\sum_{k=0,m} \exp(z_k)} \qquad (4)$$

in which

$$z_k = a_k\theta + c_k, \text{ for } k=0,m \ .$$

The number of alternatives in the multiple choice item is m. The parameters of the model are a_k and c_k, (both for k=0,m) and d_k (for k=1,m). All items considered here were four-alternative, multiple-choice items, so m=4, and there are 14 parameters in the model. [There are actually only 11 unconstrained parameters, as a single constraint is imposed in estimation on each of the sets $\underline{a}$, $\underline{c}$, and $\underline{d}$; for details of this and some interpretation of these parameters, see Thissen and Steinberg (1984).]

Throughout the remainder of this study, the "correct" value of θ that we shall try to recover in the simulations is the value given in *this model*. Similarly, all response probabilities that will be fit by the other models are generated by this model.

Aside

This model does not lend itself to easy intuitive penetration. Shortly, we shall provide some graphic evidence of its performance which usually aids understanding. Recent research (Winsberg, Thissen & Wainer, 1984) has indicated that the use of a spline function can provide the flexible fit required in this application with greater statistical stability of the parameters. This formulation seems to be of great prospective use in test construction and item analysis, although we still favor a simpler parameterization for test scoring.

II.a Model Fitting for Four Items

Why just four items? We begin this study with four items for four reasons. First, many of the properties of the estimators we have examined reveal themselves in this limited context; thus we could sharpen our intuition

If one is concerned only with correlation of ability estimates with some other variable, then one is concerned only with the linear dependence of the estimates on the true values. For each estimator, there exists a linear rescaling which minimizes the MSE of the estimates from the true values. We call this rescaling MI. The smaller the MSE under this criterion, the stronger the linear dependence of the estimates on the true values. The smaller the MSE is under this rescaling, the more the estimates would correlate with any variable linearly related to the true values.

A way to look at VM and MI together is that MI is regression of the estimates on the true values, and VM is standardization. Regression (of Y on X) can be separated into two steps (for X with "standard" variance):

1. Rescale Y to have the "standard" variance of X. After this step, use the 45-degree line through the mean as "principal axis" prediction; this is VM.

2. "Regress" the predicted values on Y to mean-Y. This gives the "regression line" instead of the principal axis. This is MI. Thus, MI is produced by "the correlation between the estimates and the true values" times VM.

V. The tests

We have examined the performance of the 28 estimators using four "tests." Two of the tests consisted of only 4 items, one was 20 items, and another was 40. The 4-item tests served two purposes: (a) they formed a basis for the construction of the longer tests, and (b) they provide evidence about the performance to be expected of the various estimators if they are used in a computerized adaptive test (CAT), in which case an estimate based on only 4 items would be required early in each testing session. The 20-and 40-item tests simulate ability estimation for realistic-length CAT and paper-and-pencil tests. Of course, all the tests and all of the item response data are artificial. The simulated item response data are probablistically determined by the trace lines for the items which make up each test.

In order to lend an air of realism to the simulated data, we created the tests using trace lines estimated from real item response data; this seemed a more realistic approach than simply "making up trace lines." The data used were from the National Opinion Research Center (NORC) national probability sample tested on ASVAB Form 8A (Bock and Mislevy, 1982). The trace lines used were generated by the parameters of the multiple category model estimated by Thissen and Steinberg (1984). Only eight distinct trace lines were used in the entire simulation; four of these were estimated from data for 4 items in the ASVAB Word Knowledge subtest and the other four from 4 items of the General Science subtest. We used these trace lines because first, they were conveniently available; second they are among the most "realistic" mathematically defined trace lines obtainable; and third, among the eight curves is represented just about any shape that the trace line for the correct response could reasonably be expected to follow. We do not claim that our simulated tests are ASVAB Word Knowledge or General Science, or any other test

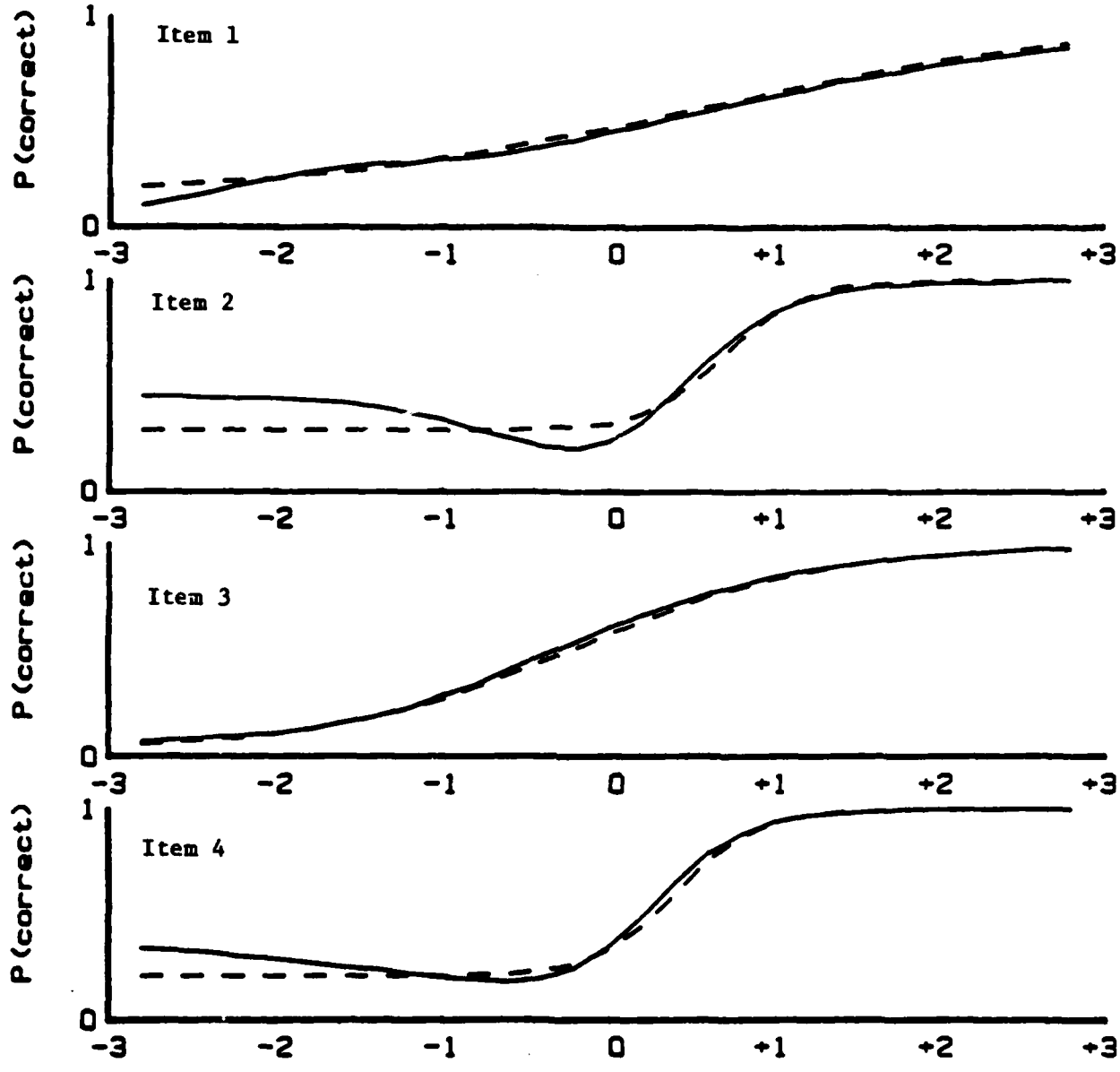
for that matter. We would claim that the simulated tests are "inspired" by trace lines that fit the item response data for those tests well, so the simulated tests are made up of items which, as best we can tell, behave like some of the items on those tests. With that disclaimer, we will continue to identify the simulated tests with "WK" and "GS" (for Word Knowledge and General Science) because those identifiers have served to keep the tests straight in our analyses and because there may be some interest in the parentage of the tests.

V.a The Word Knowledge (WK) 4-item set

The trace lines used to create the simulated data for the WK 4-item set are shown in Display 3; the parameters (for the model in equation 4) are given in Table 2. Display 3 shows the true trace lines (solid) and the best-fitting 3-PL trace lines (dashed) for comparison. The items vary in both the shapes of their trace lines and the closeness with which the binary logistic models can approximate them. WK item 1 (the items are numbered 1 to 4 here; these are not ASVAB item numbers) has an extremely modest slope and nearly linear shape; it is well approximated by any of the three logistic models at all points on the ability dimension. WK item 2 has the most "strange" (objectionable) trace line in this set; there is a very high probability (of guessing?) on the left, followed by a distinct "trough" near zero.

In the vicinity of zero on WK item 2, some of the distractors are quite effective. They fade, and a sharp rise in P(correct) appears between 0 and +1. This trace line is not well-approximated even by the 3-PL: at -2, the 3-PL underestimates the probability of a correct response by more than .1, whereas around zero it similarly overestimates the true probability. WK item 3 is the most "logistic" of the set. WK item 4 is not as non-monotonic as WK item 2, but there is still a slight trough unfittable by the 3-PL, so the 3-PL underestimates P(correct) at -2 by about .1.

In the simulation, item response data were generated from the true (solid) trace lines in Display 3, and abilities (for the 3-PL estimators) were estimated using the dashed trace lines. This is "estimating ability with the wrong model," as described above. It is not possible in this type of simulation to distinguish particular observations (item responses) as "true" or "error"; the wrongness of the model takes a more subtle form. As noted above, on WK items 2 and 4, the 3-PL underestimates the probability of a correct response at -2 by a noticeable margin. Since the data were actually generated using the probabilities given by the "true" curve, this means that more simulees at -2 will respond correctly to the second and fourth items than the 3-PL (or any-PL) model "expects"; in other words, there was an excess of 0101 item response vectors. Specifically, at -2 the true model generates 9% 0101; for the 1-PL fit for these items, that probability is 1%; for the 2-PL fit it is less than 1%; and for the 3-PL fit it is only 4%. On the other hand, the true probability of the response vector 0000 (all incorrect) is only 27%; the 1-PL fit gives 61% (!), the 2-PL fit gives 60%, and the 3-PL fit gives 39%. So no particular "0101," for instance, is an "error"; it is just



Display 3. "True trace lines" (solid) and 3-PL estimates of the four-item WK-inspired example. The fitted 3-PL curves are dashed.

Table 2. Multiple category and n-PL parameters, WK 4-item set

		Item 1			
a(k)	-1.076	-2.856	1.519	0.193	2.221
c(k)	0.118	-3.798	2.859	-1.919	2.741
d(k)		0.098	0.186	0.305	0.411
		Item 2			
a(k)	-3.026	-0.296	1.199	-0.703	2.826
c(k)	0.096	0.563	-1.612	0.874	0.078
d(k)		0.170	0.239	0.139	0.452
		Item 3			
a(k)	-0.747	-1.810	-0.069	0.687	1.938
c(k)	-1.594	-3.201	0.203	2.040	2.553
d(k)		0.170	0.239	0.139	0.452
		Item 4			
a(k)	-1.608	-0.364	-0.221	-0.774	2.967
c(k)	0.019	-1.269	0.356	0.379	0.514
d(k)		0.098	0.186	0.305	0.411

Estimated 1-PL parameters

Item	Location	Slope
1	0.10	1.04
2	0.12	1.04
3	-0.31	1.04
4	-0.02	1.04

Estimated 2-PL parameters

Item	Location	Slope
1	0.14	0.64
2	0.13	0.89
3	-0.26	1.32
4	-0.01	1.51

Estimated 3-PL parameters

Item	Location	Slope	Lower Asymptote
1	0.57	0.78	0.13
2	0.68	4.24	0.29
3	-0.20	1.39	0.03
4	0.36	4.10	0.21

that there are too many of them as far as any of the logistic models are concerned. The effect of this "excess" of patterns like 0101, and 0100, and 0001, etc. and the corresponding lack of 0000s on the distribution of estimated abilities at -2 is what the simulation is all about. With 4 items, this is somewhat comprehensible. With more items, it is not, but the idea remains the same.

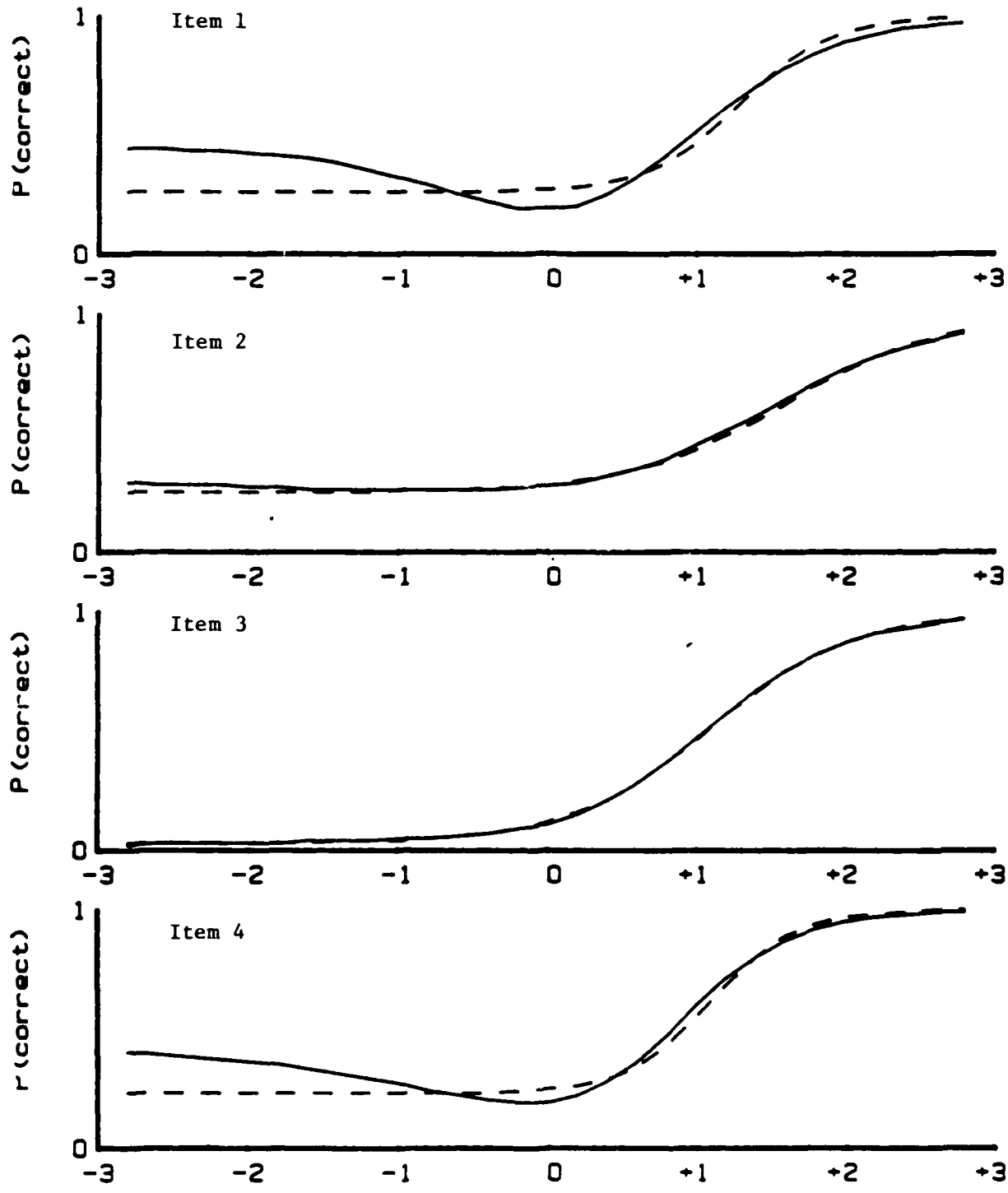
We would also note that the n-PL models are not terribly wrong. The idea has been to make them "realistically" wrong. Some of the trace lines are well approximated by some of the n-PL curves. Some are not. It is likely that the real world is like that. This is not a "worst case" simulation, but rather, an attempt at realism.

We chose the trace lines for four of the most difficult items on the Word Knowledge subtest because they are "located" farthest to the right of the distribution of the population tested. This provides more information about peculiarities of shape of the trace lines in the middle and on the left than can be obtained from data for easier items. (Correct trace lines must be uninteresting on the right, as they approach unity or something is seriously wrong.) Thus, this 4-item test provides most information between abilities of 0 and +1, near the 3-PL location estimates (in Table 2). This should be borne in mind as we discuss the results of the simulation in terms of the ability levels of our simulees (at -2, -1, 0, +1, and +2). Those ability levels should not be viewed as "absolutes," but rather relative to the simulated test. When we consider the simulees at ability level -2, we are considering examinees whose ability is 2 or more standard deviations "below" the test items to which they are responding. For this test, simulees of ability about 0 or +1 are responding to a very appropriate test, and those at +2 are responding to items which are too easy. So the outcome should be interpreted in those terms: the performance of the estimators at abilities of 0 and +1 should be indicative of their performance when items presented and the abilities of the examinees are well-matched, and performance elsewhere describes the outcome with less appropriate items.

V.b. The GS 4-item set

The trace lines used to create the simulated data for the GS 4-item set are shown in Display 4; the parameters for the multiple category model are given in Table 3. Display 4 shows the true trace lines (solid) and the best-fitting 3-PL trace lines (dashed) for com-parison. GS items 1 and 4 have troughs near 0 on the ability scale and are not fitted well by the 3-PL. GS item 2 is well approximated by the 3-PL, not by either the 1-Pl or 2-PL; GS item 3 is well approximated by all three logistic models. These 4 items are difficult, with most of their information concentrated around +1 on the ability scale.

Relative to the WK 4-item set, the GS 4-item set has more information at ability levels around +1, but the logistic models are less accurate there, especially on GS items 1 and 4. The logistic models are also very wrong at abilities below -1 on both GS item 1 and GS item 4. The slopes for the four GS items are quite high; see the n-PL parameters approximating these curves in Table 3, or inspect the trace lines. So this test is made up of very "good" items, with the exception of the peculiarities in their trace lines.



Display 4. "True trace lines" (solid) and 3-PL estimates for the four-item GS-inspired example. The fitted 3-PL curves are dashed.

Table 3. Multiple category and n-PL parameters, GS 4-item set

	Item 1				
a(k)	-2.211	-0.234	0.779	-0.278	1.943
c(k)	0.566	0.658	-1.809	1.079	-0.494
d(k)		0.185	0.185	0.185	0.445

	Item 2				
a(k)	-0.766	0.640	-0.575	-0.450	1.151
c(k)	1.741	-2.905	1.602	-0.275	-0.162
d(k)		0.183	0.183	0.183	0.451

	Item 3				
a(k)	-0.767	-0.057	-1.285	0.198	1.911
c(k)	-0.037	0.554	0.544	-0.229	-0.832
d(k)		0.264	0.264	0.264	0.208

	Item 4				
a(k)	-1.612	-0.703	-0.168	0.096	2.387
c(k)	0.475	0.451	-0.528	0.462	-0.860
d(k)		0.188	0.188	0.188	0.436

Estimated 1-PL parameters

Item	Location	Slope
1	0.93	0.73
2	1.00	0.73
3	1.90	0.73
4	0.87	0.73

Estimated 2-PL parameters

Item	Location	Slope
1	1.36	0.47
2	1.37	0.51
3	1.10	1.68
4	0.88	0.72

Estimated 3-PL parameters

Item	Location	Slope	Lower Asymptote
1	1.30	3.25	0.26
2	1.56	1.94	0.25
3	1.10	2.14	0.03
4	1.08	3.50	0.23

V.c. The GS 20-item set

In order to generate a 20-item artificial test without completing the mind-boggling excercise of somehow generating and describing 20 different trace lines, we settled upon the strategy of "multiplying" the GS 4-item set. We replicated each of the four GS trace lines five times, to make a 20-item set. The shapes remained the same as those shown in Display 4. To give the test a broader range of difficulty or, equivalently, to spread the information over a wider range of ability, each replication of the 4 items of the GS set was "translated" .5 to the left. So the first 4 items of the 20-item set have true trace lines identical to those shown in Display 4. Items 5 to 8 are the same, except that they are translated left .5, so, for instance, the trough which is at zero in GS item 1 is at -.5 in item 5. Items 9 to 12 of the 20-item set consist of the GS 4-item set translated to the left (easier) by 1, items 13 to 16 are 1.5 "easier" and items 17 to 20 are two standard deviations easier.

Since the GS 4-item set was quite difficult, this process of replicating it in successively easier versions produced a test with a good deal of information from about +2 (the top of the information in the original set) down to about -1. Table 4 gives the parameters for the multiple category model used to generate the 20-item set, as well as the n-PL parameters approximating the true trace lines. The progressive shifts of the items are more clearly visible in the n-PL location parameters (going from items 1 to 20 in sets of four) than in the multiple category parameters.

The n-PL location parameters do not shift in steps of exactly .5, because the effects of the lack of fit of the "wrong models" change as the regions in which the n-PL models do not fit "slide off" the plots to the left. So the 20-item simulation is of a reasonably broad-range test made up of highly discriminating items with slightly odd trace lines.

V.d. The GS 40-item set

The structure of the 40-item set is extremely simple; the test consists of two identical replications of the 20-item set. Each trace line in the 20-item set appears twice. Thus, the only difference between the 40-item set and the 20-item set is length. All of the trace lines are (shifted) versions of the GS curves in Display 4, and all of the parameters are the same as those in Table 4, except that there are two copies of each.

Table 4. Multiple category and n-PL parameters for 20-item set

	Item 1				
a(k)	-2.211	-0.234	0.779	-0.278	1.943
c(k)	0.566	0.658	-1.809	1.079	-0.494
d(k)		0.185	0.185	0.185	0.445

	Item 2				
a(k)	-0.766	0.640	-0.575	-0.450	1.151
c(k)	1.741	-2.905	1.602	-0.275	-0.162
d(k)		0.183	0.183	0.183	0.451

	Item 3				
a(k)	-0.767	-0.057	-1.285	0.198	1.911
c(k)	-0.037	0.554	0.544	-0.229	-0.832
d(k)		0.264	0.264	0.264	0.208

	Item 4				
a(k)	-1.612	-0.703	-0.168	0.096	2.387
c(k)	0.475	0.451	-0.528	0.462	-0.860
d(k)		0.188	0.188	0.188	0.436

	Item 5				
a(k)	-2.211	-0.234	0.779	-0.278	1.943
c(k)	-0.540	0.541	-1.419	0.940	0.478
d(k)		0.185	0.185	0.185	0.445

	Item 6				
a(k)	-0.766	0.640	-0.575	-0.450	1.151
c(k)	1.358	-2.585	1.315	-0.500	0.414
d(k)		0.183	0.183	0.183	0.451

	Item 7				
a(k)	-0.767	-0.057	-1.285	0.198	1.911
c(k)	-0.421	0.526	-0.099	-0.130	0.124
d(k)		0.264	0.264	0.264	0.208

	Item 8				
a(k)	-1.612	-0.703	-0.168	0.096	2.387
c(k)	-0.331	0.099	-0.612	0.510	0.334
d(k)		0.188	0.188	0.188	0.436

	Item 9				
a(k)	-2.211	-0.234	0.779	-0.278	1.943
c(k)	-1.645	0.424	-1.030	0.801	1.449
d(k)		0.185	0.185	0.185	0.445

	Item 10				
a(k)	-0.766	0.640	-0.575	-0.450	1.151
c(k)	0.975	-2.265	1.027	-0.725	0.989
d(k)		0.183	0.183	0.183	0.451

Table 4 (continued)

		Item 11			
a(k)	-0.767	-0.057	-1.285	0.198	1.911
c(k)	-0.804	0.497	-0.741	-0.031	1.079
d(k)		0.264	0.264	0.264	0.208
		Item 12			
a(k)	-1.612	-0.703	-0.168	0.096	2.387
c(k)	-1.137	-0.252	-0.696	0.558	1.527
d(k)		0.188	0.188	0.188	0.436
		Item 13			
a(k)	-2.211	-0.234	0.779	-0.278	1.943
c(k)	-2.750	0.307	-0.641	0.662	2.421
d(k)		0.185	0.185	0.185	0.445
		Item 14			
a(k)	-0.766	0.640	-0.575	-0.450	1.151
c(k)	0.592	-1.945	0.740	-0.950	1.565
d(k)		0.183	0.183	0.183	0.451
		Item 15			
a(k)	-0.767	-0.057	-1.285	0.198	1.911
c(k)	-1.188	0.469	-1.383	0.068	2.035
d(k)		0.264	0.264	0.264	0.208
		Item 16			
a(k)	-1.612	-0.703	-0.168	0.096	2.387
c(k)	-1.943	-0.604	-0.780	0.606	2.720
d(k)		0.188	0.188	0.188	0.436
		Item 17			
a(k)	-2.211	-0.234	0.779	-0.278	1.943
c(k)	-3.856	0.190	-0.251	0.523	3.392
d(k)		0.185	0.185	0.185	0.445
		Item 18			
a(k)	-0.766	0.640	-0.575	-0.450	1.151
c(k)	0.209	-1.625	0.452	-1.175	2.140
d(k)		0.183	0.183	0.183	0.451
		Item 19			
a(k)	-0.767	-0.057	-1.285	0.198	1.911
c(k)	-1.571	0.440	-2.026	0.167	2.990
d(k)		0.264	0.264	0.264	0.208
		Item 20			
a(k)	-1.612	-0.703	-0.168	0.096	2.387
c(k)	-2.749	-0.955	-0.864	0.654	3.914
d(k)		0.188	0.188	0.188	0.436

Table 4 (continued)

Estimated 1-PL parameters

Item	Location	Slope
1	0.63	1.27
2	0.67	1.27
3	1.28	1.27
4	0.59	1.27
5	0.31	1.27
6	0.38	1.27
7	0.68	1.27
8	0.17	1.27
9	-0.16	1.27
10	0.00	1.27
11	0.07	1.27
12	-0.36	1.27
13	-0.72	1.27
14	-0.45	1.27
15	-0.56	1.27
16	-0.97	1.27
17	-1.32	1.27
18	-0.94	1.27
19	-1.18	1.27
20	-1.63	1.27

Estimated 2-PL parameters

Item	Location	Slope
1	1.36	0.47
2	1.37	0.51
3	1.10	1.68
4	0.88	0.73
5	0.37	0.94
6	0.55	0.75
7	0.56	1.83
8	0.18	1.22
9	-0.16	1.37
10	-0.00	0.97
11	0.05	1.93
12	-0.31	1.68
13	-0.62	1.69
14	-0.47	1.16
15	-0.45	1.97
16	-0.77	2.03
17	-1.08	1.87
18	-0.92	1.31
19	-0.94	1.98
20	-1.23	2.26

Table 4 (concluded)

Estimated 3-PL parameters

Item	Location	Slope	Lower Asymptote
1	1.30	3.25	0.26
2	1.56	1.94	0.25
3	1.11	2.14	0.03
4	1.08	3.50	0.23
5	0.77	2.83	0.23
6	1.06	1.84	0.24
7	0.61	2.12	0.03
8	0.56	3.17	0.21
9	0.24	2.53	0.20
10	0.54	1.77	0.24
11	0.10	2.08	0.03
12	0.03	2.94	0.19
13	-0.32	2.32	0.17
14	0.02	1.71	0.22
15	-0.41	2.05	0.02
16	-0.50	2.78	0.17
17	-0.88	2.16	0.13
18	-0.51	1.65	0.21
19	-0.92	2.01	0.01
20	-1.03	2.66	0.15

VI. Results, natural scaling

In this section, we discuss the outcomes of the simulation study as they appear if the estimators are not rescaled. These results are important for two reasons. The first is that there may be circumstances under which there is no way to rescale, or "unshrink" estimates of ability; under such circumstances, the results of this section are indicative of the performance of the various models and estimators. The second reason for the importance of these results is that they point up sharply the need for some kind of rescaling for many of the estimators we consider. Results as they appear when the estimators are rescaled are given in Section VII.

In this and the following section, most of the results are presented in the form of XTREE glyphs, because they show the results quickly in ways that tables do not. Thissen and Wainer (1984) describe XTREE glyphs and their interpretation in this context. Briefly, XTREE glyphs are multivariate graphical displays. In the context of this investigation, each model (1-PL, 2-PL, and 3-PL) at each level of ability (-2, -1, 0, +1, +2) is assigned an XTREE; each of the 9 or 10 estimators for that model at that level of ability is assigned a pair of branches on the XTREE. The left branch has length proportional to that estimator's squared bias, while the right branch has length proportional to that estimator's random variance; so the sum of the lengths of the two branches is proportional to MSE. It sounds complicated, but it is simple: short branches are good, and long branches are bad. A standard set of branches at the bottom of each XTREE provide numerical standards for interpretation of the other branches. The assignment of estimators to branches is given by a "key" on the left side of each XTREE plot.

VI.a The 4-item simulations

Displays 5 and 6 are the XTREE plots for the outcome of the WK and GS 4-item simulations, respectively. One outcome is of such great magnitude that it dominates the scale of the XTREEs and dwarfs all other results, and that is the massive bias for most of the estimators at both extremes (-2 and +2) in both plots. The "standard" base of the XTREEs in both plots on the bias (left) side is 4, which is equal to a bias of 2 (squared); when an estimator's branch approaches that length (and some exceed it!), indicating that the mean of the estimates was about zero, the population mean, and about as far from the correct value of -2 or +2 as such estimates can be. This bias is, of course, due to "shrinkage" toward the population mean of zero. The shrinkage of estimates like these is (approximately) proportional to a ratio of the information in the prior to the information in the item responses. Most of the estimators can extract so little information from 4 items that the prior dominates the item responses, instead of the other way around.

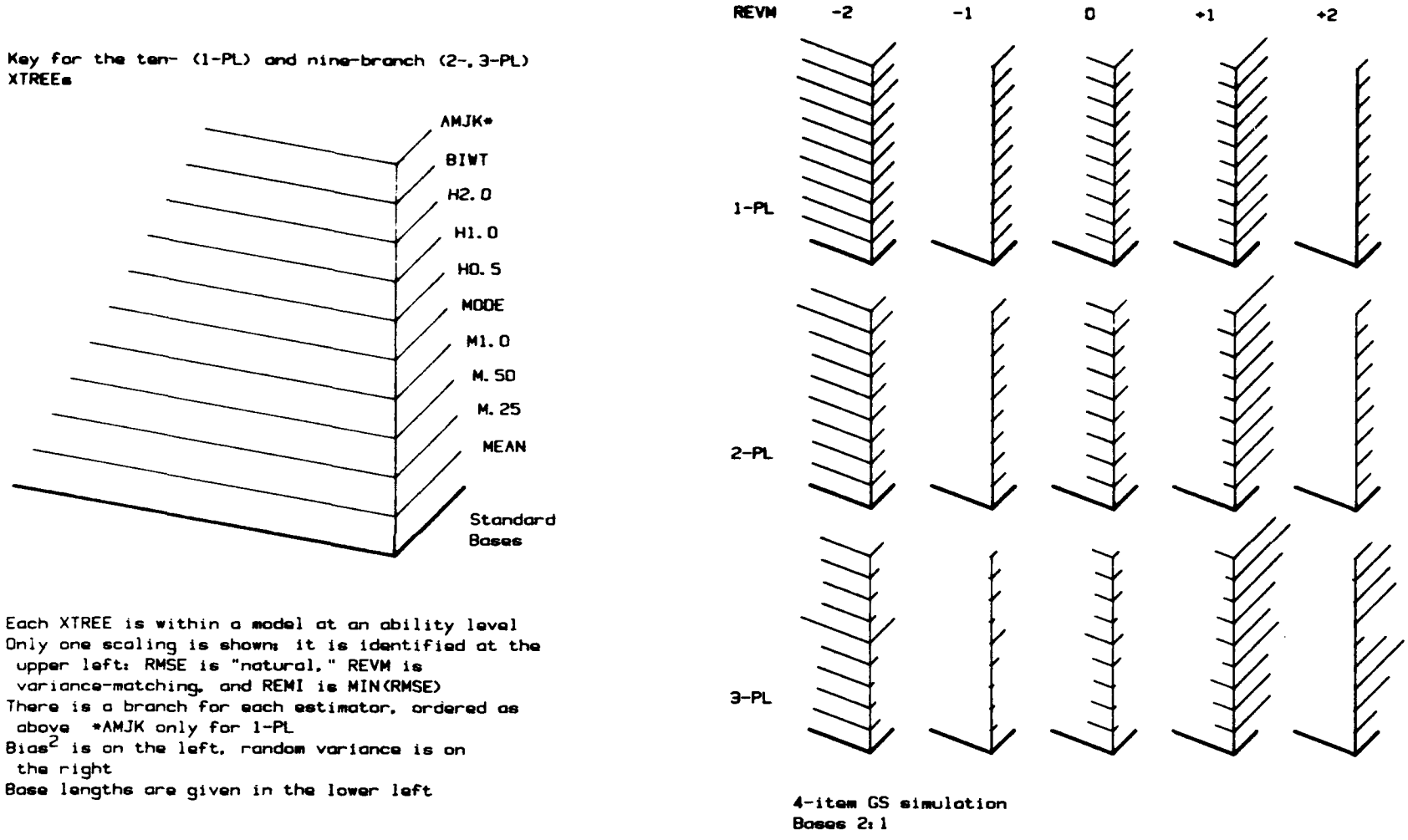
The GS 4-item set simply provides so little information at -2 that no real pattern is visible there. In the GS simulation at +2, and in the WK simulalation at both -2 and +2, there is a further pattern visible. With a

single notable exception (discussed later), the more "robustified" the estimator is, the worse it does. This effect is visible in the "scallops" on the left side of all the XTREEs in those columns, because, going from the bottom to the top, the order of the estimators is MEAN, and then three progressively robustified integrated estimators (M.25, M.50, and M1.0), then the non-robustified MODE (which does better), followed by four robustified modal estimators (H0.5, H1.0, H2.0, and BIWT). The robustified estimators "resist" some item responses they (anthropomorphically) "feel" are unreasonable, thus throwing away some of the meagre information in the 4-item data and, as a result, are shrunk (biased) even more than their conventional counterparts. There is not much to see in these XTREEs at abilities of -1, 0, and +1 because the scale is set to fit the wild behavior at the extremes on the plots. The estimators do not behave so very much differently from each other at those ability levels, although the tendency for the robust estimators to "overshrink" is still apparent.

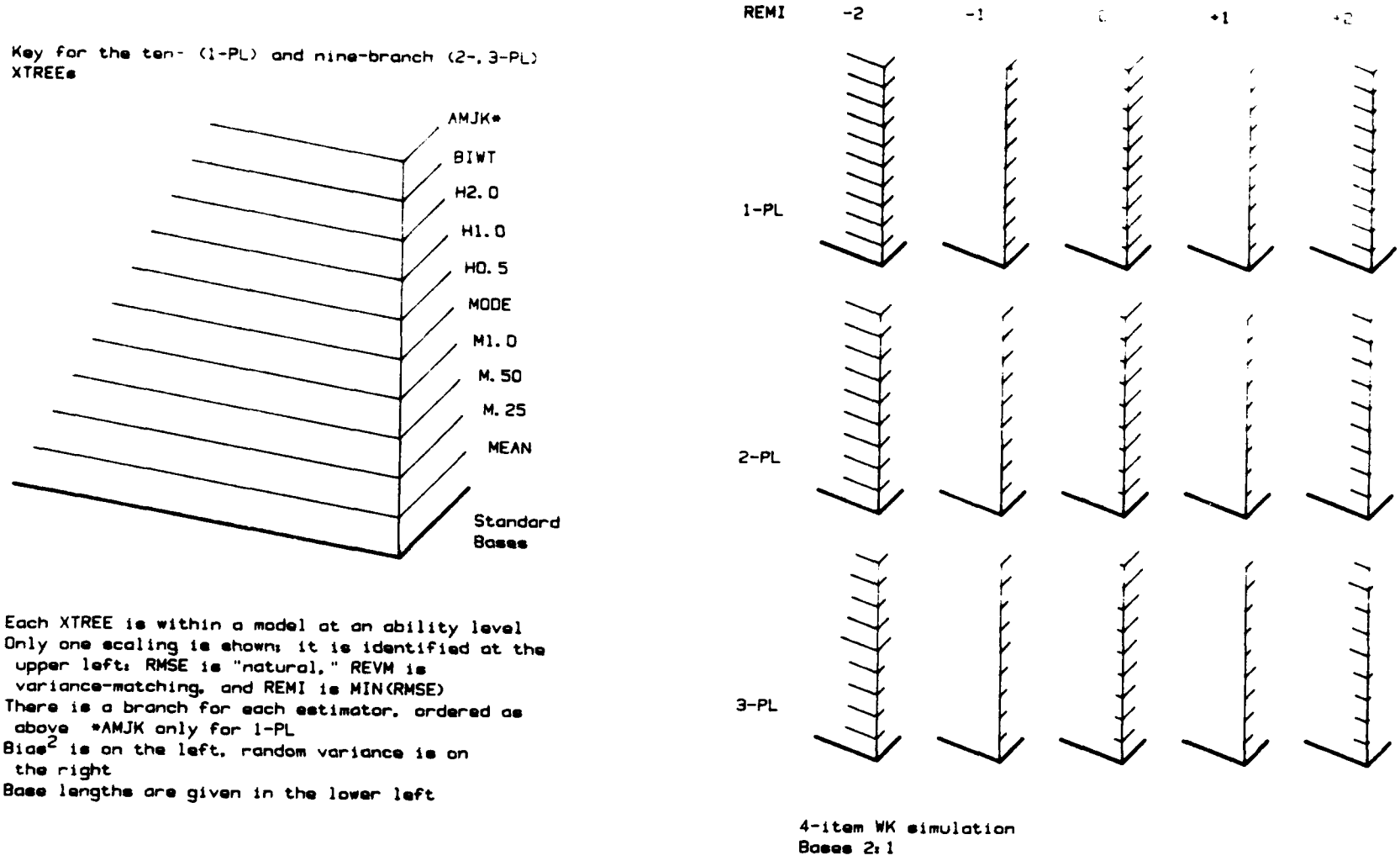
The notable exception to this generally disastrous behavior of the estimators in their attempts to reach low and high ability levels on the basis of only 4 items is AMJK. It has much less bias under almost all conditions with 4 items than do any of the other estimators. This is clearly a result of the jackknife component of the procedure. Jackknifing was originated as a procedure for reducing bias in statistical estimates (Quenouille, 1956); "shrinkage" is a kind of bias, and the jackknife eliminates some of it. AMJK has a good deal of random variance (right branches on the top trees) at all levels of ability because, since it is a robustified estimator, it, too, is throwing away some of the information in the data. But it is remarkably unbiased.

To make this more concrete, the results from the simulation program displaying bias and variance in semi-graphic form is included in Displays 7 and 8 for AMJK and the 3-PL MODE for the WK 4-item simulation. At ability = -2, 32 of the AMJK estimates are within a half a standard deviation of being right! By contrast, all of the 3-PL modal estimates are more than 1.2 away from the true value of -2.

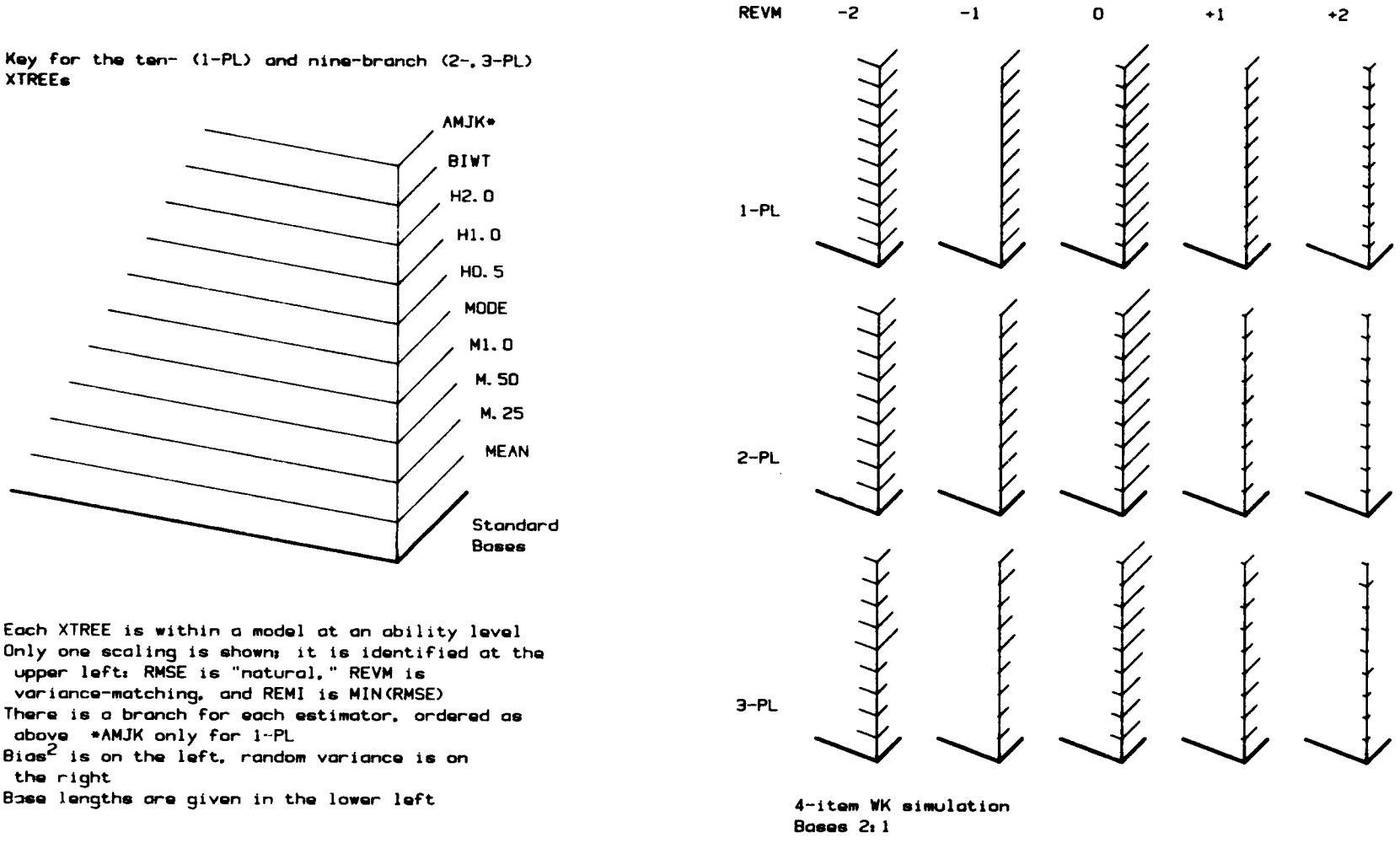
None of the IRT estimators really performed well with only 4 items, if they were trying to reach extremes of ability. (If they were trying to reach the middle, they didn't do too badly because they didn't go anywhere.) The problem is shrinkage, and some way must be found to overcome that problem before IRT estimates with small numbers of item responses can be expected to be useful. It might be noted here that "not using" a prior distribution would not solve the problem in any noncosmetic sense: if there had been no "prior," or a uniform (un-normed) prior, the many "perfect" response vectors 0000 and 1111 would have been assigned infinite estimates by all estimators. Since some of these appear at all ability levels, it would have been hard to tell about bias, but the variance of the estimates at each ability level would have been infinite. That is not better.



Display 16. XTREE Plot showing the Squared Bias and Variance for the Simulated 4 item General Science Test with REVM Rescaling.



Display 15. XTREE Plot showing the Squared Bias and Variance for the Simulated item Word Knowledge Test with REMI Rescaling.



Display 14. XTREE Plot showing Squared Bias and Variance for the Simulated 4 item Word Knowledge Test with REVM Rescaling.

Display 13 shows the 20-item results for 3-PL MODE. It is between .1 and .2 more biased at extreme abilities, but its nominal standard errors are estimated to be the same as those for the MEAN.

Additionally, it is clear from these results that with the 3-PL MEAN, the use of the normal population distribution is a non-issue at test lengths over 20 items or so. Except (slightly) at ability = -2, where this set of items provides almost no information, the item responses swamp the prior, and there is negligible bias due to shrinkage. In an adaptive test, in which the examinees of low ability (eventually) responded to some very easy items, there would probably be less bias at -2, as well.

VII. Results, with rescaling

Rescaling, either to match the variance of the generating distribution of ability or to minimize MSE, is the great equalizer of this study. The tendency for clear winners (and losers) to appear, such as existed in the previous section (AMJK (if anything) good for small sets of items and 3-PL MEAN good for long tests), disappears instantly when the estimators are rescaled in an attempt to make their various interactions with the prior vanish. But certainty will not exist in this section: Every estimator is about as good as every other (for 20 or 40 items) or as bad (for 4 items). There are some trends: pretty much without exception for the longer tests and with some exceptions for the short tests, the 3-PL estimators perform better than the corresponding 1-PL and 2-PL estimators. And the robust estimators show some advantages here, but they are by no means sufficiently uniform to recommend them for general use.

VII.a The 4-item simulations

Displays 14 and 15 show the XTREEs for the WK 4-item set with scaling to match the variance of the generating distribution; and Displays 16 and 17 give the same results for the GS 4-item set. Note that, because the badness of the estimates has been re-distributed much more evenly across levels of ability than it was originally, the bases for the XTREEs are only 2 on the bias^2 side and remain 1 on the variance side. It all fits, because a great deal of what used to be bias at the extremes has been transformed by rescaling into random variance pretty much all over. Most of the rescalings involve some mean shift (to reduce the bias at -2 which existed in all the estimators) and multiplication by a factor greater than unity to expand the variance. These processes reduced bias at the extremes, at the expense of inducing some bias and a great deal of heretofore unseen random variance in the middle.

The results with the two rescaling conventions do not differ a great deal with respect to differences among the estimators. AMJK was least changed in rescaling, of course, since it was closest in its original form. It seems that all of the estimators can be made to behave much like a transformed AMJK with an appropriate linear transformation. That is, with 4 items, they can show little bias and lots of variance. There are occasional long branches on the XTREES which attract attention, but if that estimator is followed across

4 GS X 5 Estimator=MODE

Binned u-UHAT

Bin End	u=-2.0	u=-1.0	u= 0.0	u= 1.0	u= 2.0
1.6- 1.8					
1.4- 1.6					
1.2- 1.4					
1.0- 1.2					2
0.8- 1.0					3
0.6- 0.8			2	1	16
0.4- 0.6		6	5	11	22
0.2- 0.4		16	10	18	14
0.0- 0.2		21	22	32	43
-0.2- 0.0		25	34	18	
-0.4--0.2	18	15	13	13	
-0.6--0.4	31	11	11	2	
-0.8--0.6	28	4	2	4	
-1.0--0.8	15	1	1	1	
-1.2--1.0	6	1			
-1.4--1.2	2				
-1.6--1.4					
-1.8--1.6					
End					
Mean d	-0.638	-0.082	-0.054	0.049	0.347
Std. Dev.	0.251	0.330	0.293	0.299	0.267

Binned Estimated s.e.

Bin End	u=-2.0	u=-1.0	u= 0.0	u= 1.0	u= 2.0
1.9- 1.8					
1.8- 1.7					
1.7- 1.6					
1.6- 1.5					
1.5- 1.4					
1.4- 1.3					
1.3- 1.2					
1.2- 1.1					
1.1- 1.0					
1.0- 0.9					
0.9- 0.8					
0.8- 0.7					
0.7- 0.6	10				
0.6- 0.5	35	5			
0.5- 0.4	41	29		1	43
0.4- 0.3	14	65	36	21	52
0.3- 0.2		1	64	78	5
0.2- 0.1					
0.1- 0.0					
Mean s.e.	0.492	0.387	0.300	0.298	0.388
Std. Dev.	0.076	0.057	0.011	0.023	0.061

Display 13. The frequency distributions of the Bias and Standard Error for the 3-PL MODE estimator from the Simulated 20 item General Science test.

4 GS X 10 Estimator=MEAN

Binned u-UHAT

Bin End	u=-2.0	u=-1.0	u= 0.0	u= 1.0	u= 2.0
1.6- 1.8					
1.4- 1.6					
1.2- 1.4					
1.0- 1.2		1			
0.8- 1.0		1			1
0.6- 0.8		4			2
0.4- 0.6		10	1	7	18
0.2- 0.4		18	17	19	35
0.0- 0.2	10	20	41	27	18
-0.2- 0.0	26	21	30	33	11
-0.4--0.2	20	19	10	12	15
-0.6--0.4	23	1		2	
-0.8--0.6	12	5	1		
-1.0--0.8	8				
-1.2--1.0	1				
-1.4--1.2					
-1.6--1.4					
-1.8--1.6					
End					
Mean d	-0.358	0.055	0.033	0.034	0.185
Std. Dev.	0.283	0.345	0.195	0.218	0.290

Binned Estimated s.e.

Bin End	u=-2.0	u=-1.0	u= 0.0	u= 1.0	u= 2.0
1.9- 1.8					
1.8- 1.7					
1.7- 1.6					
1.6- 1.5					
1.5- 1.4					
1.4- 1.3					
1.3- 1.2					
1.2- 1.1					
1.1- 1.0					
1.0- 0.9					
0.9- 0.8					
0.8- 0.7					
0.7- 0.6					
0.6- 0.5	6				
0.5- 0.4	75	14			15
0.4- 0.3	17	44			39
0.3- 0.2	2	42	87	90	46
0.2- 0.1			13	10	
0.1- 0.0					
Mean s.e.	0.445	0.318	0.217	0.222	0.327
Std. Dev.	0.057	0.070	0.016	0.016	0.070

Display 12. The frequency distributions of the Bias and Standard Error for the 3-PL MEAN estimator from the Simulated 40 item General Science Test.

4 GS X 5 Estimator=MEAN

Binned u-UHAT

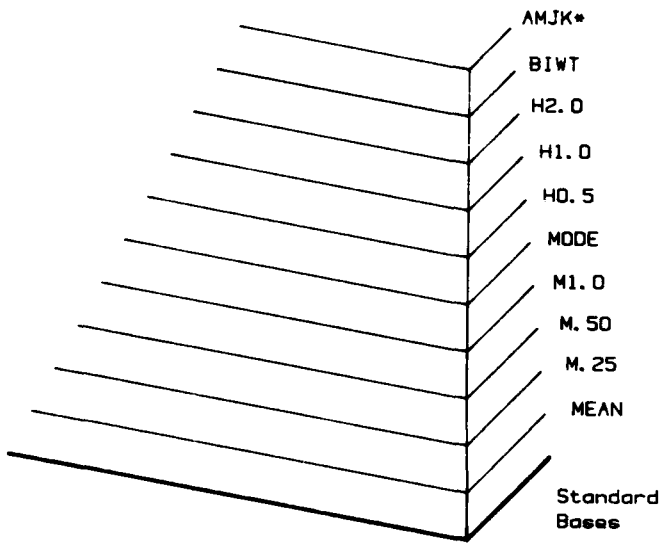
Bin End	u=-2.0	u=-1.0	u= 0.0	u= 1.0	u= 2.0
1.6- 1.8					
1.4- 1.6					
1.2- 1.4					
1.0- 1.2					1
0.8- 1.0					1
0.6- 0.8		3	1	1	19
0.4- 0.6		16	11	22	22
0.2- 0.4		22	18	20	14
0.0- 0.2		13	14	15	
-0.2- 0.0	9	15	17	11	43
-0.4--0.2	30	10	19	24	
-0.6--0.4	39	19	19	2	
-0.8--0.6	13	1	1	4	
-1.0--0.8	4	1			
-1.2--1.0	3			1	
-1.4--1.2					
-1.6--1.4	2				
-1.8--1.6					
End					
Mean d	-0.486	0.019	-0.028	0.066	0.270
Std. Dev.	0.254	0.388	0.338	0.367	0.323

Binned Estimated s.e.

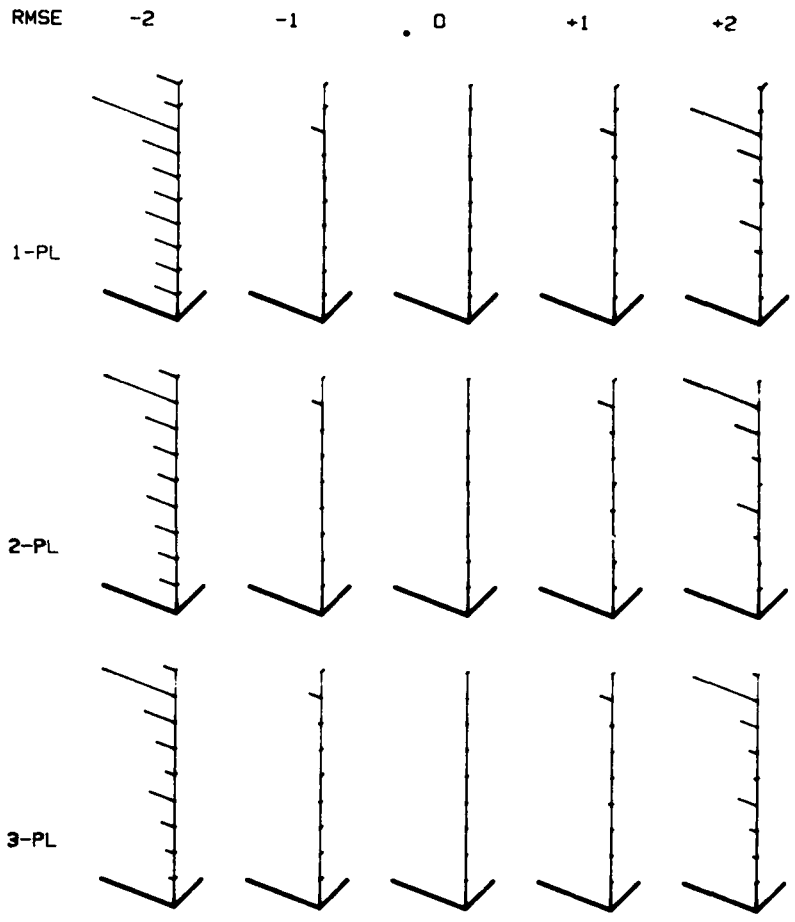
Bin End	u=-2.0	u=-1.0	u= 0.0	u= 1.0	u= 2.0
1.9- 1.8					
1.8- 1.7					
1.7- 1.6					
1.6- 1.5					
1.5- 1.4					
1.4- 1.3					
1.3- 1.2					
1.2- 1.1					
1.1- 1.0					
1.0- 0.9					
0.9- 0.8					
0.8- 0.7					
0.7- 0.6					
0.6- 0.5	55	11		1	43
0.5- 0.4	40	54	32	35	29
0.4- 0.3	5	22	25	22	4
0.3- 0.2		10	23	30	22
0.2- 0.1		3	17	10	2
0.1- 0.0			3	2	
Mean s.e.	0.497	0.413	0.312	0.328	0.427
Std. Dev.	0.047	0.091	0.109	0.104	0.111

Display 11. Frequency distributions of Bias and Standard Error for the 3-PL MEAN estimator from the Simulated 20 item General Science test.

Key for the ten- (1-PL) and nine-branch (2-, 3-PL) XTREEs

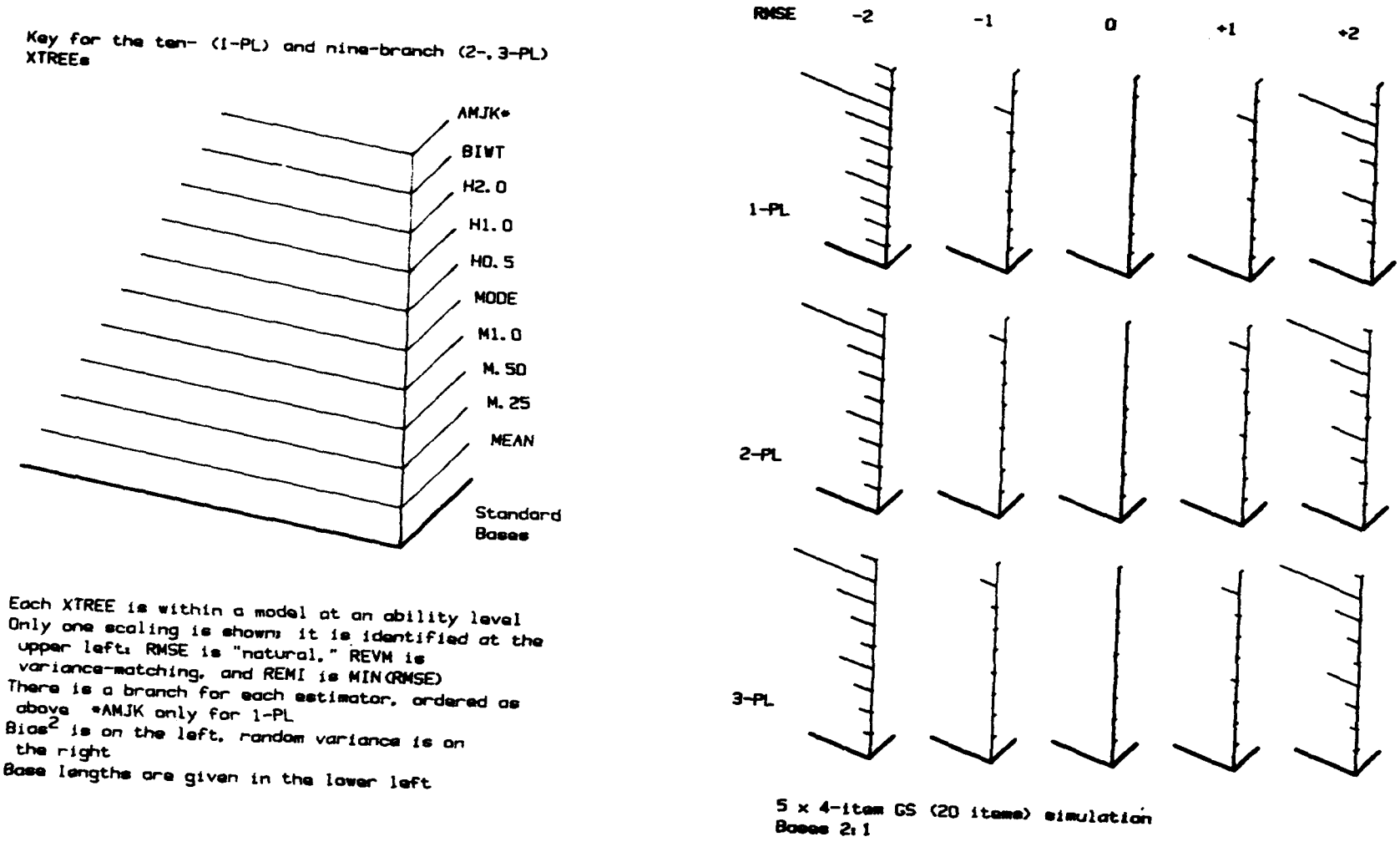


Each XTREE is within a model at an ability level
Only one scaling is shown; it is identified at the upper left: RMSE is "natural," REVM is variance-matching, and REMI is MIN(RMSE)
There is a branch for each estimator, ordered as above *AMJK only for 1-PL
Bias^2 is on the left, random variance is on the right
Base lengths are given in the lower left



10 x 4-item GS (40-item) simulation
Bases 2:1

Display 10. XTREE Plot showing Squared Bias and Variance for the Simulated 40 item General Science Test with RMSE Rescaling.



Display 9. XTREE plot showing Squared Bias and Variance for the Simulated 20 item General Science Test with RMSE rescaling.

VI.b The 20- and 40-item simulations

Displays 9 and 10 are the XTREE plots for the 20-and 40-item simulations, respectively. Note that errors (both bias and random) are smaller at these test lengths, so the standard bases of the XTREEs are smaller: two on the bias^2 (left) side and one on the random (right) side. Also note that there is a clear winner, defined by having short branches (small MSE) across the entire range of ability: 3-PL MEAN.

We knew in advance that 3-PL MEAN ought to do best as test length increased because, ignoring the fact that it is the "wrong model" for the moment (it is not all that wrong), it is asymptotically optimal. What we did not know was the location of "asymptotically." It appears to be around 20 items.

The robustified estimators continue to have problems with bias due to shrinkage at extreme true abilities. In very robust estimators, the problem is very large at 20 items and is not going away in any real sense at 40 items. BIWT is generally better than the robustified mean-estimators (M.25, M.50, and M1.0), which, in turn, are better than the h-estimators. All of those estimators do improve between 20 and 40 items; AMJK, for no reason which is clear to us, does not really improve as the test gets longer, and so loses ground steadily to the MEAN.

A surprise, in some sense, is that the MODE, which has been the classical ability estimate of IRT, does slightly but noticeably worse than the MEAN at the extremes. This is especially true for the 3-PL model. In the XTREE plots, the MODE branches are the middle branches (fifth in from either end, excluding the bases); notice that at both +2 and -2 for both the 20-and 40-item tests, the branches are longer on the bias side for the MODE than for the MEAN (bottom branches). In these regions of ability, especially with the 3-PL model, the posterior density for ability given the item responses has a distinct propensity to be skewed, and the MODE of a skewed distribution is a fragile estimator. Here we are in the odd position (for veterans of the "robust estimation campaign") of pronouncing the MEAN "more robust." But it is; at least, it is more robust than the MODE. Indeed, except that it is probably a computational monster, a median might be better yet.

Display 11 shows the frequency distribution of the differences between the estimates and the true value of ability (u-UHAT) at each level of ability, and the distributions of estimated standard errors, for the 3-PL MEAN for 20 items; Display 12 shows the same results for 40 items. In neither case is performance perfect: both show some bias at ability = -2, .5 for 20 items and .4 for 40 items; so it is not even improving rapidly there. This bias is made to appear small by squaring it on the XTREE plots. But even with some bias there, the estimated standard errors (squared) provide an excellent approximation to the real MSE at each level of ability.

4 ASVAB WK Estimator=MODE

Bin End	u=-2.0	u=-1.0	u= 0.0	u= 1.0	u= 2.0
			Binned u-UHAT		
1.6- 1.8				1	1
1.4- 1.6					1
1.2- 1.4					3
1.0- 1.2					22
0.8- 1.0					73
0.6- 0.8			18	10	
0.4- 0.6			12	5	
0.2- 0.4				8	
0.0- 0.2			33	26	
-0.2- 0.0			3	50	
-0.4--0.2		42	18		
-0.6--0.4		15	8		
-0.8--0.6		5	2		
-1.0--0.8		20	2		
-1.2--1.0		7	4		
-1.4--1.2	51	7			
-1.6--1.4	6				
-1.8--1.6	6	2			
End	37	2			
Mean d	-1.658	-0.655	0.028	0.120	0.988
Std. Dev.	0.470	0.411	0.480	0.290	0.120
			Binned Estimated s.e.		
Bin End	u=-2.0	u=-1.0	u= 0.0	u= 1.0	u= 2.0
1.9- 1.8					
1.8- 1.7					
1.7- 1.6					
1.6- 1.5					
1.5- 1.4					
1.4- 1.3					
1.3- 1.2					
1.2- 1.1					
1.1- 1.0					
1.0- 0.9					
0.9- 0.8	51	42	18	1	
0.8- 0.7	18	31	24		
0.7- 0.6	5	9	21		
0.6- 0.5	18	9	7	50	73
0.5- 0.4	8	9	30	49	27
0.4- 0.3					
0.3- 0.2					
0.2- 0.1					
0.1- 0.0					
Mean s.e.	0.723	0.728	0.626	0.487	0.504
Std. Dev.	0.134	0.123	0.151	0.056	0.036

Display 8. Frequency distributions of Bias and Standard Error for the 3-PL MODE estimator from the Simulated 4 item Word Knowledge test.

4 ASVAB WK Estimator=AMJK

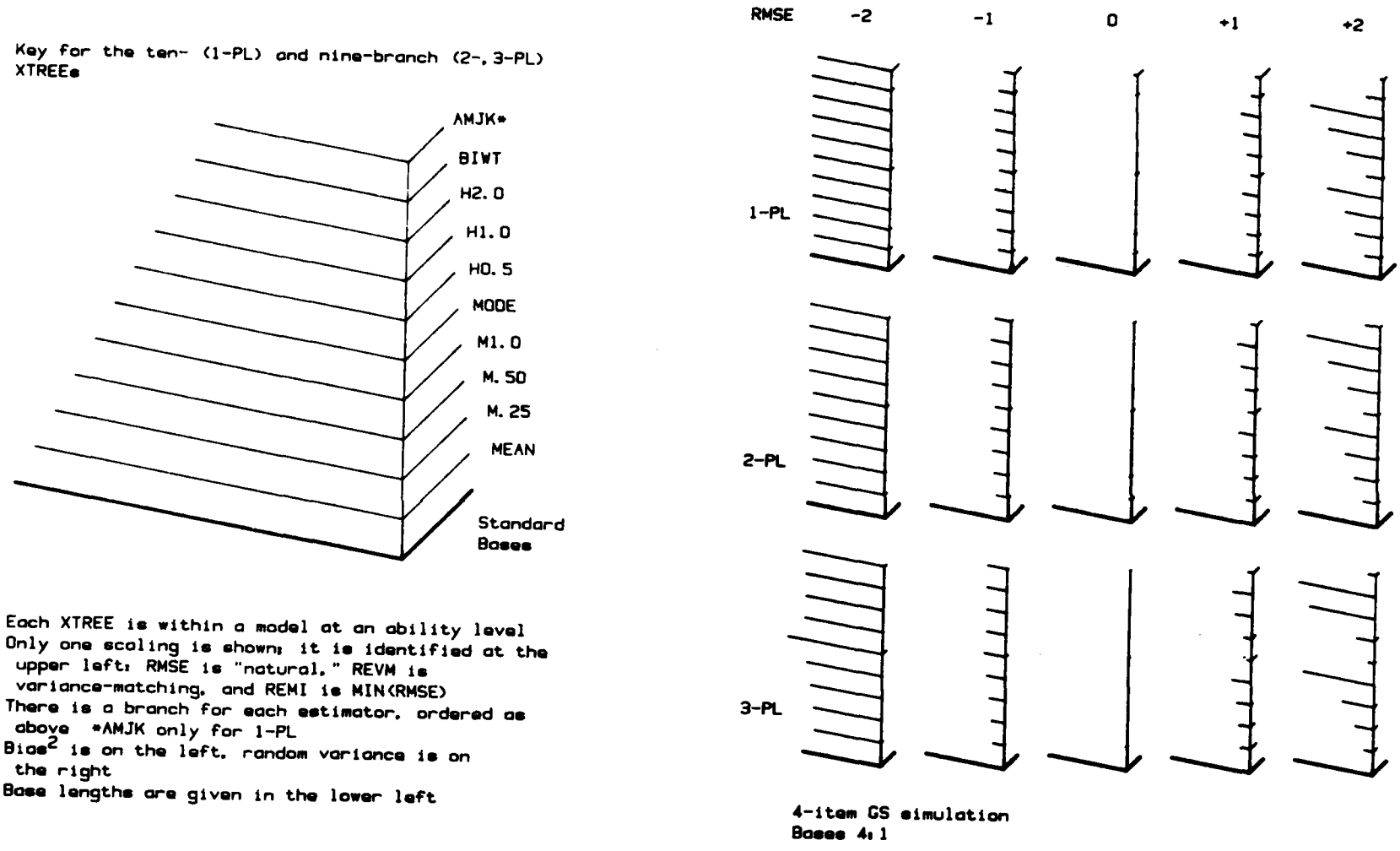
Binned u-UHAT

Bin End	u=-2.0	u=-1.0	u= 0.0	u= 1.0	u= 2.0
1.6- 1.8				1	
1.4- 1.6			15		
1.2- 1.4					26
1.0- 1.2				9	
0.8- 1.0					
0.6- 0.8			30		
0.4- 0.6		29			73
0.2- 0.4				40	
0.0- 0.2			37		
-0.2- 0.0					
-0.4--0.2		46			
-0.6--0.4	32			50	
-0.8--0.6			14		
-1.0--0.8		21			
-1.2--1.0					
-1.4--1.2	47				
-1.6--1.4			4		
-1.8--1.6		2			
End	21	2			
Mean d	-1.224	-0.274	0.288	-0.036	0.705
Std. Dev.	0.655	0.659	0.765	0.543	0.382

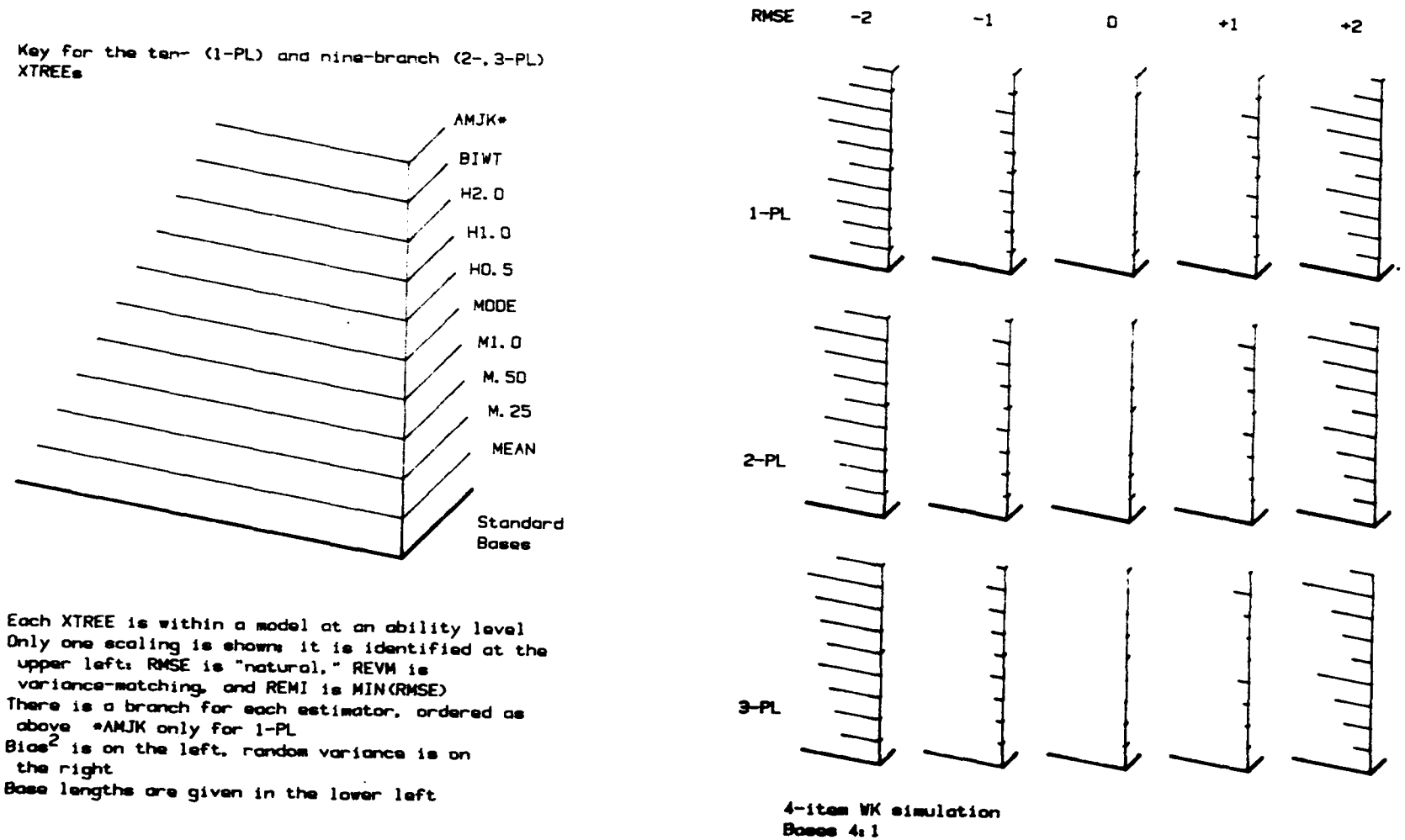
Binned Estimated s.e.

Bin End	u=-2.0	u=-1.0	u= 0.0	u= 1.0	u= 2.0
1.9- 1.8					
1.8- 1.7					
1.7- 1.6					
1.6- 1.5					
1.5- 1.4					
1.4- 1.3					
1.3- 1.2					
1.2- 1.1					
1.1- 1.0					
1.0- 0.9					
0.9- 0.8					
0.8- 0.7					
0.7- 0.6					
0.6- 0.5	12	11	4	2	
0.5- 0.4	54	58	77	48	27
0.4- 0.3					
0.3- 0.2					
0.2- 0.1					
0.1- 0.0	34	31	19	50	73
Mean s.e.	0.327	0.338	0.382	0.245	0.147
Std. Dev.	0.206	0.200	0.167	0.209	0.182

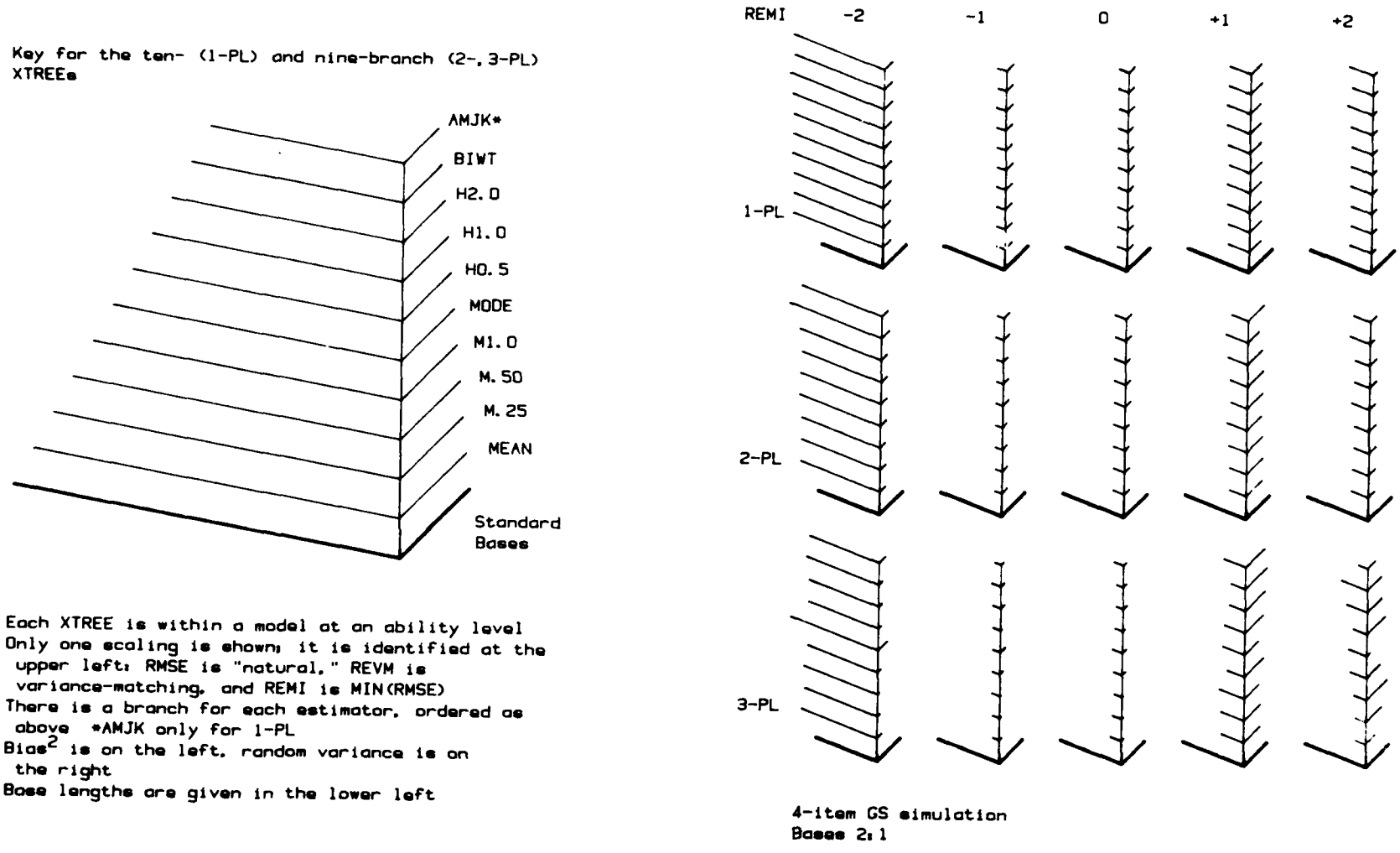
Display 7. Frequency Distributions of Bias and Standard Errors for the AMJK Estimator from the Simulated 4 item Word Knowledge test.



Display 6. XTREE plot showing Squared Bias and Variance for the Simulated 4 item General Science Test with RMSE rescaling.



Display 5. XTREE plot showing Squared Bias and Variance for the Simulated 4 item Word Knowledge Test with RMSE rescaling.



Display 17. XTREE Plot showing the squared Bias and Variance for the Simulated 4 item General Science Test with REMI Rescaling.

to other levels of ability there will be some equally surprising short branch somewhere. In most cases in the four plots there appears to be a tendency for the 3-PL estimators to do better than the others.

It may be that any procedure which succeeds in getting rid of bias at the extremes of ability with only 4 items must induce a great deal of random variance everywhere. It is not clear which is better: to administer 4 items and still believe everyone has approximately ability zero (most "naturally scaled" estimators) or to believe that everyone is all over the place, with variances sometimes in excess of the original one (for some rescaled estimators). If rescaling is available, it seems that all of the estimators are approximately equally good (which is bad).

VII.b The 20- and 40-item Simulations

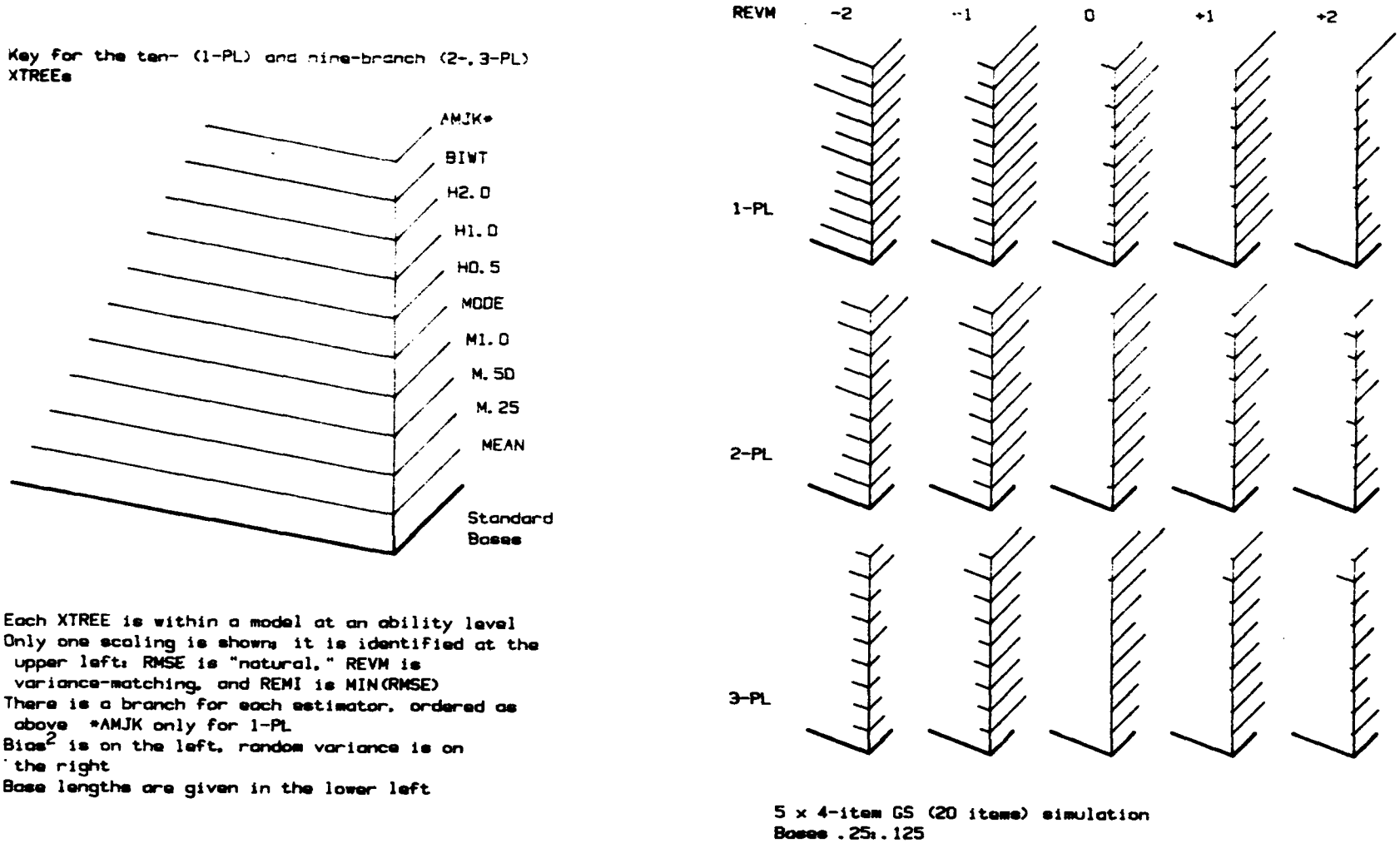
Displays 18, 19, 20, and 21 show the 20-item variance-matched and MIN(MSE) results, followed by the 40-item VM and MI results. Performance here, in contrast to the 4-item case, is so good that the scale of the XTREEs has been changed dramatically; the bases are .25 on the bias^2 side and .125 on the random variance side. So what seem to be long lines are nothing. The XTREEs are drawn to that scale to show the clear patterns, and there are some. One clear pattern in all four plots is that the 3-PL estimators perform much better than the others. So we restrict our consideration to 3-PL estimators.

The MEAN is no longer a clear winner if the estimates can be "unshrunk." In both rescalings, BIWT, H0.5, and H1.0 frequently do better (have shorter branches on both sides of the trunk) than the MEAN. M0.25 and M.50 seem to "trade in" some random error for more bias than the MEAN shows, but perform well. The MODE continues to show erratic performance, and H2.0 seems to have a good deal of bias at times, like at +2 and -2 in the 20-item plots of the results with MSE minimized.

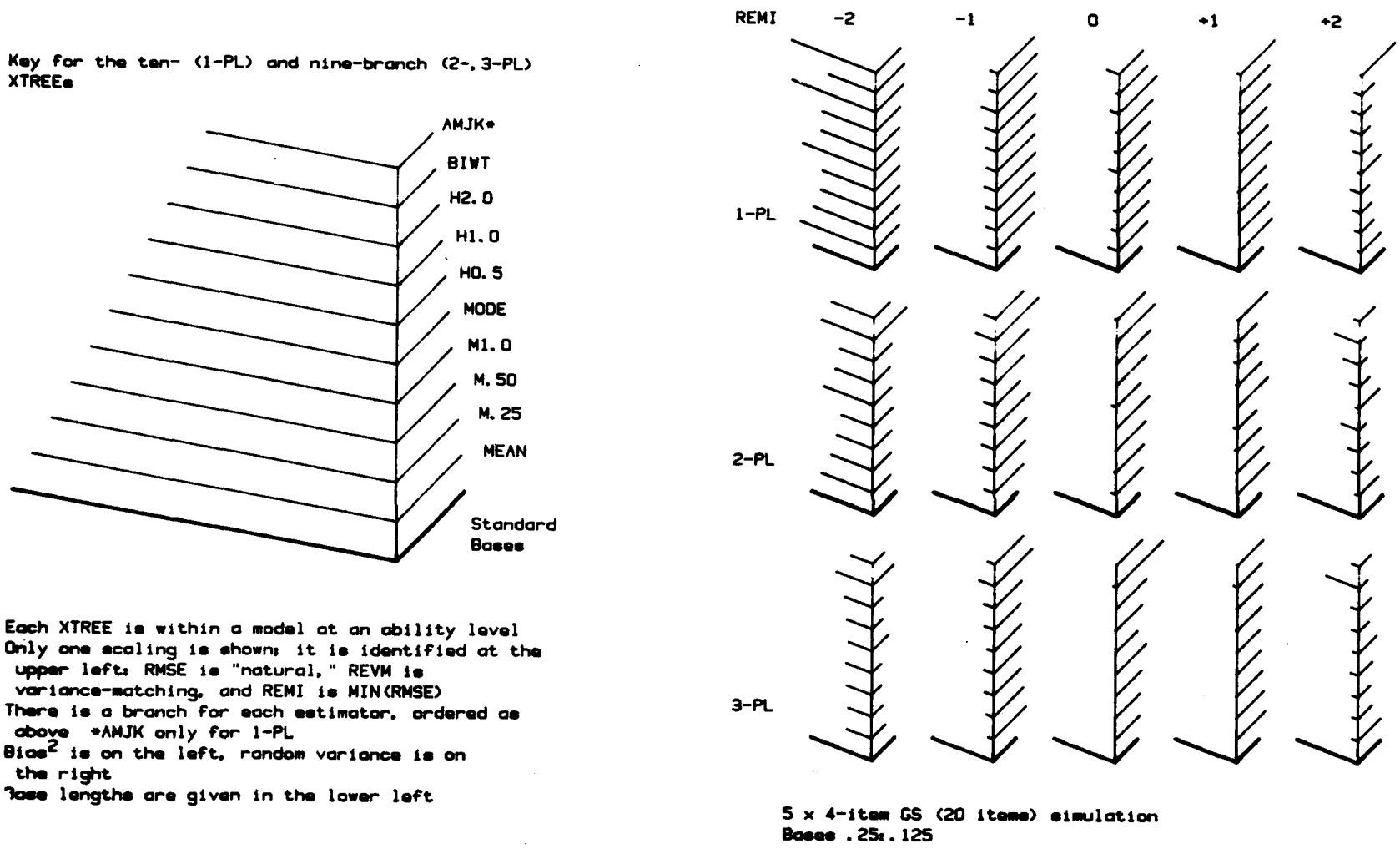
The robust estimators induce shrinkage, which yields bias. If this shrinkage can be reduced in a practical situation, robust estimators may be useful: specifically H0.5, H1.0, and BIWT. But they must be stretched to beat the MEAN on MSE. Of course, this indicates they will beat the MEAN on correlation under any circumstances. Thus, robustness can help us; but its inward regression when used in estimation with priors is a problem that must be considered in any context in which robust estimators are used.

VIII. Conclusions

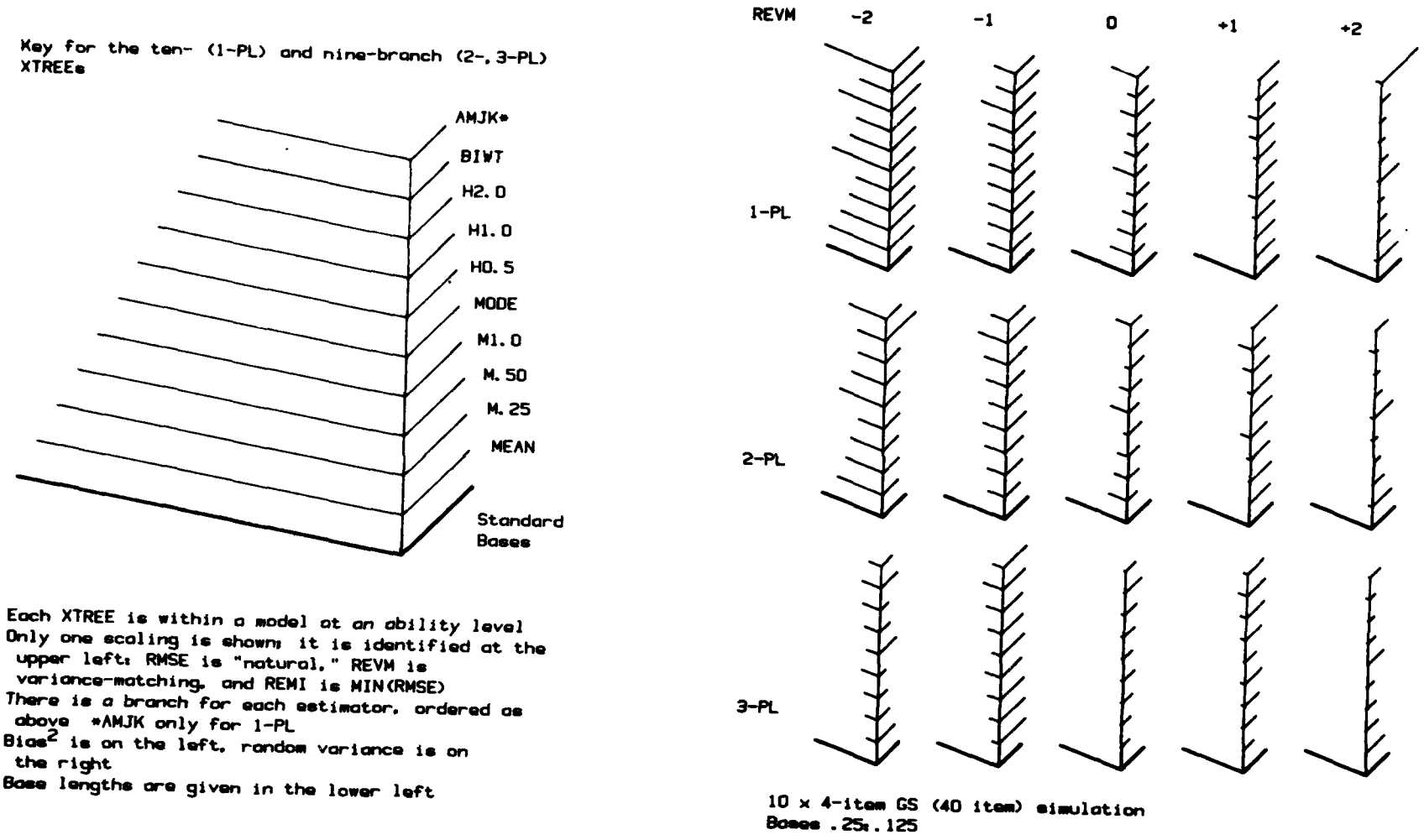
The conclusions reached in this investigation are different for two different classes of situations. The first class of situations includes those in which the numerical value of the ability estimate must be taken seriously, without rescaling. Such situations include comparing the ability estimate to some cut-point or criterion for a classification decision and also the process of computerized adaptive testing in which the value of the current ability estimate at any point in the process is used to select the next item. [In



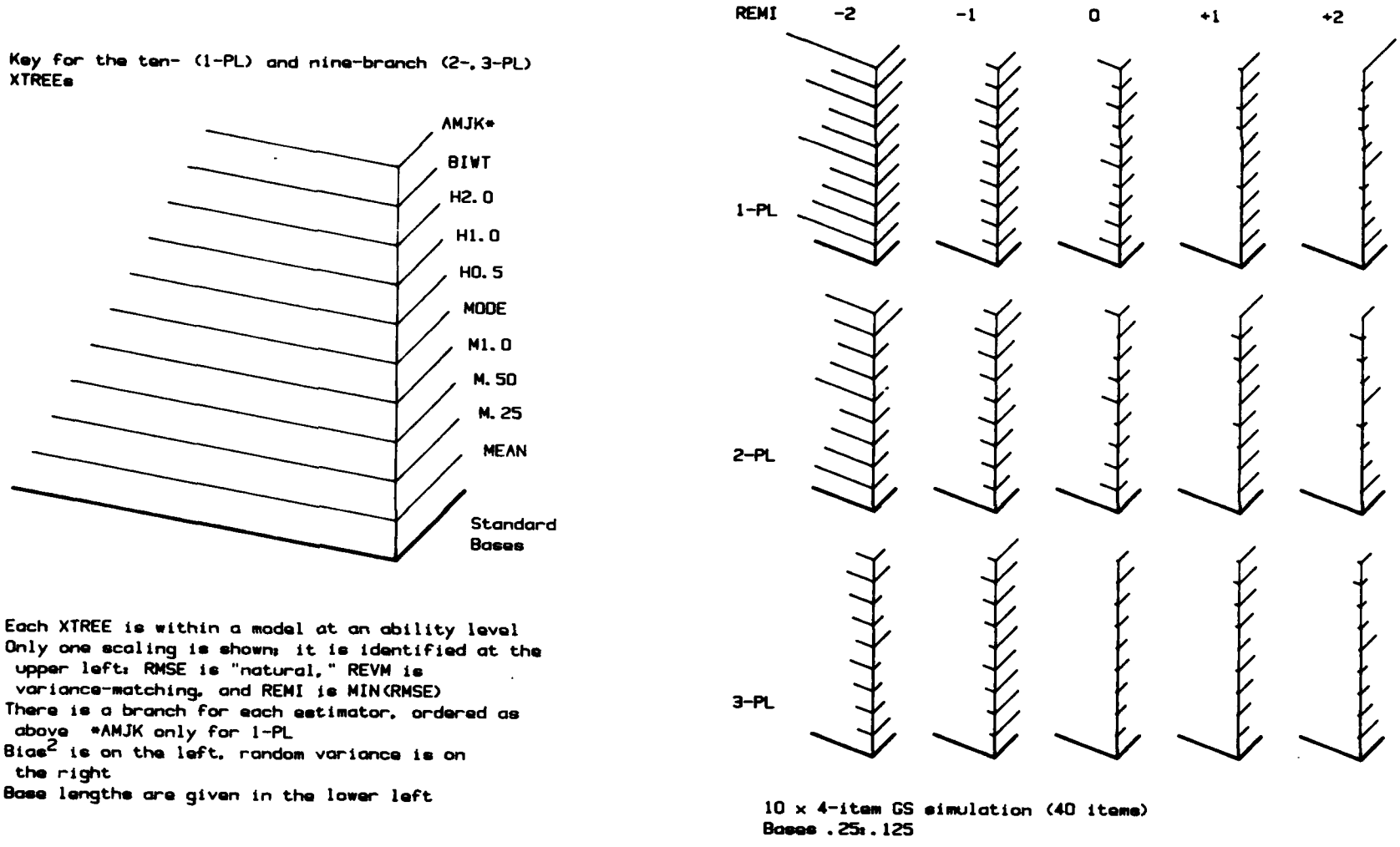
Display 18. XTREE Plot showing the Squared Bias and Variance for the Simulated 20 item General Science Test with REVM Rescaling (Note Bases are .25: .125).



Display 19. XTREE Plot showing the Squared Bias and Variance for the simulated 20 item General Science Test with REMI Rescaling.



Display 20. XTREE Plot Showing the Squared Bias and Variance for the Simulated 40 item General Science Test with REVM Rescaling.



Display 21. XTREE Plot showing the Squared Bias and Variance for the simulated 40 item General Science Test with REMI Rescaling.

The results for the naturally scaled estimators differ, depending on whether a few items (four or so) or many items (20 or 40) are considered. If the goal is to make an ability estimate with a few items, there is a serious problem with almost all of the estimators, and that is bias due to shrinkage. For a population with a mean of zero, that means that almost all of the estimates for almost all response patterns are very near zero. So the estimators do well for examinees with true ability near zero: their estimates are near zero. The estimators do not do so well with examinees with true abilities near -2 or +2, as their estimates are also near zero. With a few items (about four), most of the estimators move so little that it is not clear that the outcome is worth the computation. With items like the ones in these simulations, with substantial non-zero levels of P(correct) on the left, examinees with true abilities of -2 have estimates which tend not to move down from zero at all. The estimators do better on the right: examinees with true abilities at +2 may be assigned estimates a little over +1 by the best of the conventional estimators, 3-PL MEAN.

AMJK is a notable exception to this trend toward zero. At some cost in increased variance, AMJK "spreads" the high and low ability examinees' estimates much sooner (in number of items) than any of the other estimators. This is not a matter of being "robust"; it is a matter of being much less biased, due to the jackknife component of the procedure.

An unexpected, but nevertheless important, finding in this investigation is that, with few items, the problem with IRT ability estimators is not robustness, or lack of it, but rather, shrinkage and its avoidance. One of the robust estimators, AMJK, happened to have been constructed in such a way that it avoids shrinkage well; this lets us see the problem. Future work in ability estimation, especially for practical applications of adaptive testing, is required on "unshrinking" ability estimates. It is possible that we can do much better. It would also be useful to do much better, as shrinkage is a problem that will remain. A frequently proposed solution to the problem of shrinkage, to "not use a prior" or "use a uniform prior," does not really solve the problem in a useful way. The response vectors whose estimates shrink most are the "perfect" ones: 0000 and 1111 for 4 items. With "no prior," these response vectors either have infinite estimates of ability or are pronounced to have "no estimate." While having an excessively regressed estimate may be bad, it is not clear that either an infinite one or none at all is better. To ignore the perfect response vectors (which are quite common with a few items) and to note that the rest of the estimates are less biased if no prior is used is solving only part of the problem at the extreme expense of another part. The problem is not solved here, but it is well specified.

With many items, 3-PL MEAN does as well as its estimated standard errors say it does, and that is pretty well. This optimal estimator does well as long as there are enough item responses that "the data swamp the prior." Unless some effort is made to reduce shrinkage, the robust estimators seem to have too little information relative to the prior and to regress more from the extremes, thus losing to 3-PL MEAN.

In the second class of situations, in which the values of the ability estimates are not used with respect to outside criteria (like cut-points or the point of maximum information of pre-calibrated items) or in which rescaling is possible, there is a completely different set of conclusions. These conclusions apply to situations in which rescaling is possible, or in research applications in which only the correlation of the estimates with other variables is required, as correlation is not affected by any linear rescaling. When rescaled, all of the estimators behave similarly when faced with 4 items. They do not behave very well, as most of the rescaling increases the random variance in the process of "stretching" the scale out to the extremes.

As the test gets longer (to 20 and 40 items in our simulations), a pattern of superior performance for the 3-PL estimators emerges. This superiority of the 3-PL estimators, of course, is based on "error-free" estimates of the item parameters, which may be difficult to obtain in practice or else require large calibrating samples. And Jones, Wainer and Kaplan (1984) describe the scale of problems induced in ability estimates by error in 3-PL item parameter estimates. However, if the 3-PL item parameter estimates are good and the test is long, several 3-PL ability estimators are more-or-less equally good: 3-PL MEAN, H0.5, H1.0, and BIWT. Those robust estimators, when their excessive shrinkage is "defeated" by rescaling, perform a little better than the MEAN. This means that, even if they were not rescaled, if they were used only in a correlational context, the robust estimators would do better than the MEAN. So they are to be recommended for some applications. The other robust estimators either do not improve well as the test increases in length (AMJK and H2.0) or seem to decrease random error at too much cost in bias (the Mnnn estimators).

In conclusion, it should be observed that these simulations were realistic rather than abusive. This shows in the results: the robust estimators did not do better than the "asymptotically best if the model fits" MEAN in MSE and did very little better if the estimates were rescaled. The true trace lines we used were not very different from the n-PL trace lines which approximated them when estimating ability with the wrong model. Presumably, if we had made the wrong model "wronger," by using more exotic true trace lines that the logistic models could not approximate well, the performance of the MEAN would have deteriorated and the robust estimates would have gained on it. But it is not clear what would have been gained conceptually by such a "worst case" set of simulations.

When the model may be "moderately wrong," as it was in these simulations, prescriptions are fairly clear. With a few items, either do not compute IRT ability estimates at all until you have more items, or use AMJK. If you have more items, use 3-PL MEAN if you cannot rescale or "unshrink" or use H0.5, H1.0, or BIWT if you can rescale, or are concerned only with correlation and want some "robustness" or protection against aberrant responses.

References

Allen, M., & Yen, W. (1979). Introduction to measurement theory. Monterrey, CA: Brooks/Cole.

Andersen, E.A. (1980). Discrete statistical models with social science applications. Amsterdam: North Holland Publishing.

Bock, R.D. (1983). The discrete Bayesian. In H. Wainer & S. Messick (Eds.), In principals of modern psychological measurement. Lawrence Erlbaum Associates. Hillsdale, NJ: pp. 103-115.

Bock, R.D. & Mislevy, R.J. (1982). Applications of EAP estimation in computerized adaptive testing. Paper presented at the annual meeting of the Psychometric Society, Montreal, Canada.

Jones, D.H. (1982a). Tools of robustness for item response theory (Technical Report 82-36), Princeton, NJ: Educational Testing Service Program Statistics.

Jones, D.H. (1982b). Redescending M-type estimators of latent ability (Technical Report 82-30), Princeton, NJ: Educational Testing Service Program Statistics.

Jones, D.H., Wainer, H. & Kaplan, B. (1984). Estimating ability with three item response models when the models are wrong and their parameters are inaccurate (Technical Report 84-46), Princeton, NJ: Educational Testing Service.

Lord, F. (1980). Applications of item response theory to practical problems. Hillsdale, NJ: Lawrence Erlbaum Associates.

Lord, F. (1984). Technical problems arising in parameter estimation. Paper presented at the 1984 meeting of the American Educational Research Association: Washington, DC.

Lord, F. & Novick M.R. (1968). Statistical theories of mental test scores. Reading, MA: Addison-Wesley.

Mislevy, R.J., & Bock, R.D. (1982). Biweight estimates of latent ability. Educational and Psychological Measurement, 42, 725-737.

Quenoille, M.H. (1956). Notes on bias in estimation. Biometrika, 43, 353-360.

Rasch, G. (1960). Probabilistic models for some intelligence and attainment tests. Copenhagen: Danmarks Paedagogiske Institut.

Thissen, D., & Wainer H. (1982) Some standard errors in item response theory. Psychometrika, 47, 397-412.

Thissen, D., & Steinberg, L. (1984) A response model for multiple choice items. Psychometrika, 49, 501-520.

Thissen D., Wainer, H., & Rubin, D. (1984). A computer program for simulationn evaluation of IRT ability estimators (Technical Report 84-50). Princeton, NJ: Educational Testing Service.

Thissen, D., & Wainer, H. (1984). The graphical display of simulation results, with applications to the comparison of robust IRT estimators of ability. (Technical Report 84-49) Princeton, NJ: Educational Testing Service.

Wainer, H., & Wright, B. (1980). Robust estimation of ability in the Rasch model. Psychometrika, 45, 373-390.

Winsberg, S., Thissen, D., & Wainer, H. (1984). Fitting item characteristic curves with spline functions (Technical Report 84-52). Princeton, NJ: Educational Testing Service.

Wright, B., & Stone, M. (1979). Best test design, Chicago, IL: MESA Press.

Glossary

A number of mnemonics are used throughout the series of reports on the robust IRT estimation of ability (here, and those by Thissen, Wainer and Rubin (1984) and Thissen and Wainer (1984)). Some of these are standard in item response theory and some are less so as the structure of the study requires us to make distinctions in nomenclature not commonly of interest. The names we use, mostly four capitalized characteristics in length for the convenience of some computer systems, are brought together here with verbal descriptions; more mathematical descriptions are provided in various places in the text.

Ability

The central concept in the study is "ability"; the standard notation for this latent variable in IRT is θ. We use θ frequently in the text, but also U (both upper and lower case) in some of the material which is (essentially) computer output, for the simple reason that most computer output devices cannot print θ.

Models

1-PL This nearly-standard nomenclature is used for the one-, two-,

2-PL and three-parameter logistic IRT models. The models are defined

3-PL in equations (1), (2), and (3); the 3-PL model is that of Lord (1980), the 2-PL has all lower asymptote parameters at zero, and the 1-PL model also has all the slopes equal.

Estimators

MEAN The MEAN (also called "EAP" by Bock and Mislevy (1982)) is the average, obtained by numerical integration, of the posterior density over θ, given the item responses.

Mnnn There are three "robustified" mean-type estimators considered in the study, denoted M.25, M.5, and M1.0. These are the expected value analogs of the Jones (1982a) "h-estimators" in which the contribution of each item to the posterior is weighted by an exponential function of the information it provides about θ in that region. The values of nnn are the values of "h" used.

MODE The MODE, more commonly called the "Bayes modal" estimate in the IRT literature because it includes the population density, is the mode of the posterior density over θ given the item responses.

Hnnn Three examples of the class of "robustified" modal estimators proposed by Jones (1982a) are included: H0.5, H1.0, and H2.0. These are modal estimators in which the contribution of each item to the posterior is weighted by an exponential function of its information at that point. The number following "H" is the value of Jones' (1982a) "h", the exponent of the information.

BIWT Another modal estimator in which the contribution of each item is weighted, this time with Tukey's "biweight," was proposed by Bock and Mislevy (1982).

AMJK Wainer and Wright (1980) proposed a relatively complex robust estimator of ability which is a weighted jackknifed modal estimator.

With the exception of AMJK, all estimators are defined for all models, so an estimator is defined by both its four-character identification and that for the model with which it is used.

Scaling Conventions

Estimates of θ are determined only up to an arbitrary linear transformation (of location and scale). For some purposed, i.e., comparing scores for a single group of examinees who all responded to the same set of test items, the scale is irrelevant as between-individual comparisons have the same properties under any linear transformation. For other purposes, i.e., comparing the values of different estimators which may be on different scales on bias or MSE, the choice of linear transformation of the estimates is crucial. We consider three scaling conventions for all of the estimators.

No special name is given to the "natural scaling of the estimators, defined as "the way they come out of the computer program." This Scale is defined by an interaction between information in the item set and the variance of the population distribution (which is always 1 here), and so varies from estimator to estimator as each estimator behaves as though there was a different amount of information in the item responses.

VARM One procedure to make the performance of the various estimators more straightforwardly comparable is to rescale all estimators to have the same variance in the simulee sample, as well as the same mean. By rescaling the estimates obtained with each estimator to have the same

mean (zero) and the same variance (2) as the uniform (-2,2) sample of simulees used in this study, we make the concept of MSE much more meaningful.

MINV — If MSE is to be the criterion for selection of a good estimator, another rescaling of the estimates obtained with each estimator is that which minimizes MSE over the sample of simulees.

Performance Criteria

Several variables describe the performance of each estimator in the simulation study. Each of these variables is defined for each scaling convention (above).

UHAT — A vector of simulated values for each estimator at each level of θ. The simulees are not indexed; functions of this vector [i.e., S.D.(UHAT)] are over the simulees.

UBAR — The average of the estimates for each estimator at a particular level of θ (or u or U) is called UBAR. These values are different for the three scaling conventions, and so UBAR is modified by the name of a scaling convention (none, VARM, or MINV).

DBAR, DBVM, DBMI — The bias of the estimates for an estimator is the difference between UBAR for that estimator and the true value of θ. This is denoted DBAR for "natural" scaling, DBVM for VARM rescaling, and DBMI for MINV rescaling.

S.D., SDVM, SDMI — The standard deviation of the estimates for a particular estimator at some level of θ is denoted by S.D. for the "natural" scaling, SDVM for VARM rescaling, and SDMI for MINV rescaling.

RMSE, REVM, REMI — The Root Mean Square Error is the root of the sum of the squares of DBAR and S.D. (mutatis mutandis for the other scaling conventions) and is denoted in forms parallel to the bias and standard deviation indices.

☆U.S. GOVERNMENT PRINTING OFFICE: 1985 569 053 20008

END

FILMED

6-85

DTIC